Department of Computer Science
University of Sheffield

On the design, implementation and experimental evaluation of a novel gateway architecture for the GSM Short Message Service

Guillaume Peersman

July 2002
Submitted for the degree of Doctor of Philosophy.
Department of Computer Science
The University of Sheffield

Dédié à ma femme Lucy et ma fille Isabelle Rose,

Dedicated to my partner Lucy and my daughter Isabelle Rose, No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or any other institution of learning.

Mr Guillaume Peersman
04 August 2003

Table Of Contents

GPRS	General Packet Radio Service
GSM	Global System for Mobile communications

H

HLR	Home Location Register
HSDC	High Speed Data Circuit

I

IBM	International Business Machines
IMEI	International Mobile Equipment Identity
IMSI	International Mobile Subscriber Identity
ISDN	Integrated Services Digital Network
ISO	International Standard Organisation
ISUP	Isdn User Part (for signalling SS7)
ITU	International Telecommunication Union

L

LAN	Local Area Network
LAPDm	Link Access Protocol on the Dm channel

M

MAP	Mobile Application Part
MBCS	Multi-Byte character set
MIME	Multi-purpose Internet Mail Extension
MoU	Memorandum of Understanding
MM	Mobile Management
MMS	Multimedia Messaging Service
MS	Mobile Station
MSC	Mobile Switching Centre
MSISDN	Mobile Station ISDN number
MSRN	Mobile Station Roaming Number
MTP	Message Transfer Part
MTP	Message Translation Part

N

NACP	Network Access Control Part
NDA	Non Disclosure Agreement
NDL	Network Definition Language
NUA	Network User Address

O

OEM	Original Equipment Manufacturer
OFTEL	OFfice of TELecommunications

Abbreviations and Acronyms

A

AC	Authentication Centre
ACK	ACKnoledgement
AGCH	Associated Control CHannel
AIM	Application Interface Module
API	Application Protocol Interface
ART	Authorite de Regulation des Telecommunications
ASCII	American Standard Code for Information Interchange
ASN1	Abstract Syntax Notation 1

B

BCCH	Broadcast Control CHannel
BSC	Base Station Controller
BSS	Base Station System
BTS	Base Transceiver Station 1

C

CCCH	Common Control CHannel
CCITT	Comité Consultatif International Télégraphique et Téléphonique
CDMA	Code Division Multiple Access
CEPT	Conférence Européene des Postes et Télécommunications
CIMD	Computer Interface to Message Distribution

D

DBCS	Double Byte Character Set
DCS 1800	Digital Cellular System
DECT	Digital European Cordless Telecommunication
DTX	Discontinuous Transmission

E

EIR	Equipment Identity Register
E-MAIL	Electronic Mail
EMS	Enhanced Message Service
ETSI	European Telecommunication Standard Issue

F

FDMA	Frequency Division Multiple Access
FPLMTS	Future Public Land Mobile Telecommunication System

G

OSI	Open System Interconnection
OSI RM	OSI Reference Model

P

PCH	Paging CHannel
PCIA	Personal Communications Industry Association
PCM	Pulse Code Modulation
PDA	Personal Digital Assistant
PIN	Personal Identification Number
PLMN	Public Lands Mobile Network
PSTN	Public Switch Telephone Network

R

RACH	Random Access CHannel
RFC	Request For Comment
RPE-LPC	Regular Pulse Excited – Linear Predictive Coding
RR	Radio Resource

S

SACCH	Slow Associated Control CHannel
SBCS	Single-Byte Character Set
SCCP	Signalling Connection Control Part
SDCCH	Stand-alone Dedicated Control CHannel
SIM	Subscriber's Indentity Module
SME	Subscriber Mobile Entity
SMPP	Short Message Peer to Peer
SMS	Short Message Service
SMSC	Short Message Service Center
SMSG	Short Message Service Gateway
SMS-GMSC	SMS Gateway Mobile Switching Service
SMS-IWMSC	SMS Inter-Working Mobile Switching Service
SNMP	Simple Network Management Protocol
SPL	Self Provision License
SS7	Signalling System number 7

T

TAP	Telecator Alphanumeric Protocol
TAPI	Telephony API
TCAP	Transaction Capability Application Part
TCH	Traffic CHannel
TDMA	Time Division Multiple Access
TP-CD	Transfer Protocol Command Data
TP-CDL	Transfer Protocol Command Data Length
TP-CT	Transfer Protocol Command Type
TP-DA	Transfer Protocol Destination Address

TP-DCS	Transfer Protocol Data Coding Scheme
TP-DT	Transfer Protocol Data Type
TP-DU	Transfer Protocol Data Unit
TP-FCS	Transfer Protocol Failure CauSe
TP-MMS	Transfer Protocol More Messages to Send
TP-MN	Transfer Protocol Message Number
TP-MR	Transfer Protocol Message Reference
TP-MTI	Transfer Protocol Message Type Indicator
TP-OA	Transfer Protocol Originator Address
TP-PID	Transfer Protocol Protocol IDentifier
TP-RA	Transfer Protocol Recipient Address
TP-RD	Transfer Protocol Reject Duplicates
TP-RP	Transfer Protocol Reply Path
TP-SCTS	Transfer Protocol Service Centre Time Stamp
TP-SRI	Transfer Protocol Status Report Indication
TP-SRQ	Transfer Protocol Status Report Qualifier
TP-SRR	Transfer Protocol Status Report Request
TP-ST	Transfer Protocol STatus
TP-UD	Transfer Protocol User Data
TP-UDL	Transfer Protocol User Data Length
TP-VP	Transfer Protocol Validity Period
TP-VPF	Transfer Protocol Validity Period Format
TUP	Telephone User Part for SS7

U

UCP	Universal Computer Protocol
UK	United Kingdom
UMTS	Universal Mobile Telephony System
USA	United States of America

V

VLR	Visitor Location Register
VLSI	Very Large Scale Integration
VMSC	Visited Mobile Switching Center

List of figures

List of Tables

Abstract

Congestion and capacity problems of the existing mobile communication networks of the late eighties resulted in the demand for a brand new mobile telephony standard. Competition amongst the various existing standards implementation was fierce and lead to the availability of a plethora of incompatible networks and very little hope for the establishment of a global technology geared up to the expectancy of the users. Paging network were widely used as a cheaper alternative to voice enabled networks and were offering users the ability to receive textual information while on the move. Bridging the gap between the paging world and the mobile communication world was essential. The Global System for Mobile communications was designed as the European answer. The European alternative was promising a feature rich, secure and truly global system with the added ability to handle two-way paging like functionalities with the Short Message Service (SMS). The success of a new service such as SMS relies heavily on its adoption by a majority of users, which in turn is mainly a consequence of the availability of software application gateways. In the mid nineties a Sheffield based software company realising the market needs teamed up with the University's research department to produce the first commercially available gateway architecture for the SMS. The resulting work is described in this thesis.

The first part of the analysis examines the architecture of a GSM network and the building blocks of the SMS. The technical implementation is described and the fundamental properties examined such as roaming, routing, protocol limitations, usability and interoperability problems. The specification and design of the gateway architecture is then addressed with an emphasis on character set conversion, routing and queueing issues. The implementation details are then examined with a description of each of the modules. The performance of the gateway is examined with the implementation of a test bed fed by traffic generated by customers. The issues examined were: identification of bottlenecks, protocol efficiencies and an analysis of the chosen queueing model implementation.

The second part of the analysis presents the results obtained from the measurements taken for a period of a year. An analytical model was formulated to validate the results from the measurements. The comparison revealed the ability of the model to simulate the behaviour of the gateway under medium to heavy loads and highlighted the areas that would be most affected by optimisation. The important factors limiting throughput and quality of service were discovered in the capacity of the connections to network operators and policy chosen for the message queues. An alternative queueing discipline is proposed that would lead to increased fairness offered to the wide variety of applications connected to the gateway and the network operators through a single link of known capacity. Interactive conversations such as quizzes and games based are offered low latency while more bandwidth demanding ones such as mass voting applications benefit from very high throughput. The overall gateway architecture described was the first one of its kind and consequently each of its module was designed and implemented from scratch. As a result Dialogue Communications benefited from the novelty and head start needed in a fierce competitive market to position itself as a market leader in mobile applications and services, competing with global companies such as Logica, Ericcson or Microsoft.

1. Introduction

This document presents the work carried out during the past five years while reading for my PhD dating back to October 1996. At the time, digital mobile phones were starting to become affordable and no longer reserved to the corporate market. As the popularity of GSM quickly grew, users started to realise that there was more to GSM than plain voice services. In particular with the Short Message Service (SMS), users of GSM networks are able to exchange alphanumeric messages (up to 160- characters) with other users or services, almost anywhere in the world, within seconds of submission. Although the limitation in the size of the message may seem like a problem it is actually one of the advantages of the Short Message Service. Information has to be delivered efficiently and still many applications only require a small amount of bandwidth.

Some possible use for the Short Message Service include:

- *Multicasting value added services* will provide real-time delivery of stock market sheets, sports results, lottery numbers, etc. to the users of the digital cellular networks.

- *Search and retrieval of information:* Using existing standardised protocols such as Z39.50, the user will be offered the possibility to locate a computer program, a book etc. with the relevant piece of information delivered directly to his phone.

- *Real time notification services:* as more and more users spend time out of their office they still want to be able to be notified about new e-mail, voicemail, or even a fax.

The possibilities are only limited by the imagination of the developers and new applications appear daily. While text messaging quickly gained popularity, it became clear that new applications would only appear if the right kind of gateway was in place. No such gateway was available in 1996 and there was a need to both provide for an easier way to submit messages and a common interface to access the various parameters offered by SMS. This was no easy task, no framework was available to expand on and very little had been done to properly unleash the potential of text messaging in mobile network.
Starting with a broad specification, the gateway described in this document was implemented during the past 5 years and evolved from a simple converter of emails to a robust messaging platform offering easy and full access to the SMS potential through a common powerful API.
The first part of this analysis examines the technology behind GSM and SMS and highlights the most important features. The GSM standard itself is an enormous piece of work spanning tens of thousands of pages with a few thousands for SMS only. With no provision for a standard way of accessing SMSC, manufactures have developed a plethora of protocols all incompatible and offering various levels of support for the full set of features available.

The second part of the analysis introduces the fundamentals of the design of protocol gateways with an emphasis on character set conversions and a short description of one of the access mechanism supported by the gateway, namely SMTP.
The third part of the analysis focuses on the design and implementation of the gateway itself. The modular layered design is described as the reader is taken through the path followed by messages from the API access, through to authentication, message formatting, message routing and finally submission to the SMSC. The gateway acts as both a protocol converter and a complex message router with advanced capabilities for failsafe and clustering operation. The performance of the gateway is then assessed through benchmarking in a common basic configuration and an analytical model formulated to validate the results of the measurements.
There are still ongoing developments in the mobile messaging markets and SMS is set to follow the evolution of other messaging applications such as email or instant messaging. As market acceptance grows towards a commodity, users will expect the ability to send enhanced content including still and moving images, interlaced with sounds and rich text. Later, video and audio streaming technology will be adapted and implemented as new value added services to SMS.
I have been privileged to follow the evolution of SMS pretty much from the start and work with Dialogue Communications Ltd a local company who sponsored my research. During the past five years. Dialogue has grown from a two people operation to recently treble its workforce and be recognised as the leader in SMS and WAP applications in the UK and is currently sponsors two other PhD students as part of his research labs. While commercial sensitivity has limited publications to tutorial papers, the commercial implementations of the gateway both as a service run by Dialogue and other large corporate have been enormously successful. The gateway has been selected as a messaging platform for companies such as Yellow pages, BT Openworld, GUS home shopping or Nortel Networks. On the other hand, thousands of users currently subscribe to the service hosted by Dialogue and running on a cluster of the same gateway. The architecture was recently chosen by the BBC as part of a national live program on the popularity of SMS: "The joy of text", during which the cluster saw tens of thousands of messages processed.

1.1. Thesis structure

The structure of this thesis is as follows:

*Chapter 2:*introduces the Global System for Mobile Communications or GSM and the standard bodies that designed the overall architecture. A brief overview of the mobile terminal available today is also given. A detailed description of the GSM network and protocol architecture together with an explanation of the different services available follows.

Chapter 3: focuses on the Short Message Service itself with a description of the protocol architecture, and the short message submission methods. The section also raises the important question of the different alphabets used in computer systems and the problems encountered when submitting a short message to a different country.

Chapter 4 and 5: introduces computer code pages and the Simple Mail Transfer protocol

Chapter 6 and 7: Describes the work that has been carried on in the past five year and depicts the gateway software architecture. The section also presents an experimental evaluation of the gateway in a basic configuration and formulates an analytical model.

Chapter 8: addresses scalabilities issues and provides highlights on future work.

1.2. Acknowledgements

I especially wish to thank Hugh Spear and Paul Griffiths for sponsoring my work and now employing me as technical manager for Dialogue Communications Ltd. Their help, advice and guidance in the past few years have been invaluable in giving me an insight on the issues I should be focusing on. I also wish to thank Robin Keech, senior software developer at Dialogue for taking over most of the development of the gateway in the past year and immensely contributing to discussions and design decisions. I am indebted to my academic supervisor, Professor Srba Cvetkovic, for his constant moral support and guidance both while he was at Sheffield University and since he has left. I also wish to thanks Dr Gordon Manson for taking over supervision and helping me through the last stages of my PhD. Finally, a special thank to my partner Lucy Sharland for supporting my efforts throughout my studies.

2. The Global System for Mobile Communications

2.1. Introduction

The development of GSM started in 1982 when the European Conference of Posts and Telecommunications Administrations, consisting then of the telecommunications administrations of 26 nations. They established a team with the title "Groupe Spéciale Mobile", hence the term GSM witch today stands for Global System for Mobile Communications [Dechaux *et al*. 1993]. The aim of the team was to develop a common standard for a future pan-European cellular network.

The CEPT made these decisions in an attempt to solve the problem created by the uncoordinated development of individual national mobile communications systems using incompatible standards. The impossibility of using the same terminal in different countries while travelling across Europe was one of these problems.

By 1986 it was clear that most of the analogue networks would run out of capacity by the early 1990s. As a result, a directive was issued for two blocks of frequencies in the 900Mhz to be reserved absolutely for the future pan-European service to be opened in 1991. The GSM members also took the decision at that time to adopt a digital rather than analogue system.

The digital system would offer spectrum efficiency, better voice quality and new services with enhanced features including security. It would also permit the use of Very Large Scale Integration technology (VLSI) which would lead to smaller and cheaper handsets. Last but not least a digital approach would complement the development of Integrated Services Digital Network (ISDN [J.M. Griffiths, 1992]) with which GSM would have to interface.

The development of the specification and the deployment of the network is undertaken in different phases, mutually compatible and bringing more flexibility in terms of facilities. Phase 1 started in 1988 and specified the overall architecture of the network as well as the main facilities.

The European Telecommunications Standards Institute [ETSI] (see section 2.2.3) took over the group in 1989 and finalised the GSM standard in 1990. The first operational service started in 1991 and GSM has since become an international standard. So successful is GSM that many countries throughout the world have adopted the standard and GSM is now being used not only at the original 900Mhz-frequency band but also at 1800Mhz and even 1900Mhz in the USA. The main feature of GSM is the provision of good speech quality over a whole range of operating conditions, the support for international roaming and the ability to offer many new value-added services such as voice mail, call handling facilities, Call Line Identification - and the Short Message Service [ETSI 3.40 1996].

According to the GSM Association, the current statistics for GSM are:

- No. of Countries/Areas with GSM System (October 2001) - 172
- GSM Total Subscribers - 590.3 million (to end of September 2001)

- World Subscriber Growth - 800.4 million (to end of July 2001)
- SMS messages sent per month - 23 Billion (to end of September 2001)
- SMS forecast to end December 2001 - 30 Billion per month
- GSM accounts for 70.7% of the World's digital market and 64.6% of the World's wireless market

2.2. Standards bodies and associations

2.2.1. ITU

Headquartered in Geneva, Switzerland, The International Telecommunication Union (ITU) is an international organisation within which governments and the private sector co-ordinate global telecom networks and services. Its membership includes 188 countries and more than 450 public and private companies and organisations with an interest in telecommunications, information technology and broadcasting. The area of work of the ITU is in the domains of standardisation and international frequency allocation and numbering plan. (http://www.itu.int).

2.2.2. GSM MoU

The GSM MoU (Memorandum of Understanding) Association is the principle body responsible for promoting and evolving the GSM cellular radio platform worldwide. As from September 1997, the Association has a total of 256 members from 110 territories. Membership of the GSM MoU Association is currently open to licensed mobile network operators committed to building and implementing GSM based systems and government regulators/administrations who issue commercial mobile telecommunications licences. This includes GSM 900, GSM 1800, GSM 1900 and other future GSM derivatives such as mobile satellite systems utilising the GSM platform. The association is also a valuable source of information on the features offered by any GSM network in the world. (http://www.gsmworld.com).

2.2.3. ETSI

ETSI (European Telecommunication Standard Institute) Special Mobile Groupe (SMG) develops standards for the GSM (Global System for Mobile Communications) family of public digital mobile communications systems with a built-in capability for unrestricted world-wide roaming of users and/or terminals between any networks belonging to this family. Eleven SMG have been created (see appendix for details) to handle the different aspects of this extremely complex system. The SMG4 is the one responsible for the ongoing work being carried out on the data and telematic services of GSM, DCS 1800 and UMTS, including the inter-working with other networks. (http://www.etsi.fr).

The following work items are currently progressed or maintained within SMG 4:

- High Speed Circuit Switched Data (HSCSD): Higher data rates through multiple timeslots.

- Shared Inter-working Function (SIWF): Centralised functions for protocol conversion towards the fixed network for circuit switched data services (GSM 03.54)
- GPRS (General Packet Radio Service) inter-working.
- Short Message Service (SMS): Store and forward service for messages with up to 160 characters (GSM 03.40)
- Cell broadcast Short Message Service: Broadcast of messages to all mobile stations within a radio cell.
- Mobile Station Application Execution Environment (MAEE): Advanced platform for terminal based telecommunication services.
- New Multiplexing Protocol between Terminal Equipment and Mobile Station (Draft GSM 07.10)
- Mobile AT Commands and GSM API

2.2.4. PCIA

The Personal Communications Industry (PCIA) is the leading international trade association representing the personal communications services (PCS) industry. PCIA has been instrumental in advancing regulatory policies, legislation, and technical standards that have helped launch the age of personal communications services. One of PCIA's greatest strengths is its ability to foster and represent consensus in order to advance the interests of the PCS industry. PCIA represents the chief providers of wireless voice and data communications to both consumers and businesses. PCIA's member companies include PCS licensees and those in the cellular, paging, ESMR, SMR, mobile data, cable, computer, manufacturing, and local and inter-exchange sectors of the industry, as well as technicians, wireless systems integrators, communications site owners, distributors and service professionals, and private corporate system users. (http://www.pcia.com).

2.2.5. OFTEL

OFTEL (OFfice of TELecommunications) is the regulator – or "watchdog" – for the UK telecommunications industry. The organisation was set up under the Telecommunications Act 1984, and monitors and enforces the conditions in all telecommunications licences in the UK. OFTEL also initiates modifications to these licence conditions.

All telecommunications operators, local cable companies, mobile network operators and the increasing number of new operators – must have an operating licence. These set out what the operator can – or must – do or not do. The OFTEL controls the prices of the main network services and establishes milestones the local cable companies and mobile network operators must reach in building their networks.

Users of the services supplied by the operators also need a licence. In nearly all cases they are covered by a class licence – a licence issued to a group, not an individual, allowing certain activities. For example, the Self-Provision Licence (SPL) enables customers to use telephones in their homes.

OFTEL is also responsible for the management of the UK numbering plan and all numbers allocation. (http://oftel.gov.uk).

2.2.6. ART

ART (Authorite de regulation des telecommunications). Is the French equivalent to the OFTEL in the UK and performs similar tasks.(http://www.art-telecom.fr).

2.3. Mobile terminal summary

2.3.1. Basic GSM phones functionality

Most of today's mobile phones offer the same basic set of functionality, 99 memories to store phone numbers and names (depending on SIM card storage), last calls statistics (number dialed, communication duration etc.), a dozen different ringing tones and battery life of a few hours (depending on the capacity). The most sophisticated phones also support the network dependent services described in the following sections.

2.3.1.1. Caller display

Also called Caller ID, is the availability for a mobile phone to display the phone number of the person calling you; moreover, if its number is stored in the phone's memory, the name will be displayed instead. On the other hand, when making a call from a mobile phone, Caller Id will send the phone number. The availability of this option, however, depends on the caller wishing to have his phone number being sent or the originating network implementing this facility.

2.3.1.2. Call waiting

This is a network service that advises the user of an incoming call whilst being on a voice or data call, thus providing the user with the near-availability of a second phone line.

2.3.1.3. Call divert

Using this facility a user of a mobile cellular network is able to divert all the incoming calls to another phone number without his phone ringing. Different levels of diverting can be used in order to achieve the desired quality of service: Divert when busy; divert when not reachable or even when not answered.

2.3.1.4. Short message service

As described in section 3.

2.3.1.5. Fax and Data

Fax and data services are also available depending on your contract and the cellular networks. The maximum bandwidth on GSM networks is still 9600 bps (without compression). However, new faster services up to ISDN rates will be available in the near future.

2.4. Architecture of a GSM network

This section describes the architecture and various building blocks of a typical GSM network. It is based on the rather excellent if slightly dated tutorial from John Scourias [Scourias 1994], with a few additions and amendment. The layout of a generic GSM network with its several functional entities is shown on the figure below:

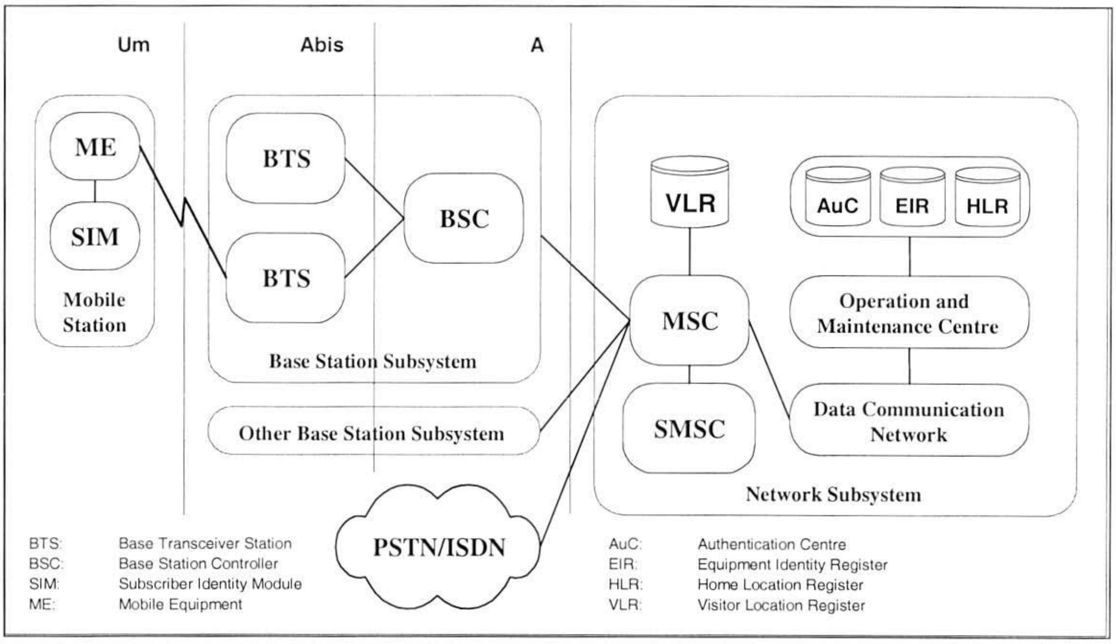

Um
Abis
A
ME
SIM
Mobile Station
BTS
BTS
BSC
Base Station Subsystem
Other Base Station Subsystem
PSTN/ISDN
VLR
AuC
EIR
HLR
MSC
SMSC
Operation and Maintenance Centre
Data Communication Network
Network Subsystem
BTS: Base Transceiver Station
BSC: Base Station Controller
SIM: Subscriber Identity Module
ME: Mobile Equipment
AuC: Authentication Centre
EIR: Equipment Identity Register
HLR: Home Location Register
VLR: Visitor Location Register

Figure 1: Architecture of a GSM network

The architecture can be divided in three main components. The Mobile Station (MS), namely the GSM terminal, is held by the subscriber, the Base Station Subsystem (BSS) controls the radio link with the Mobile Station, and the Network Subsystem performs the switching of calls and other management tasks such as authentication. The Mobile Station and the Base Station Subsystem communicate across the interface called Um, also known as the air interface or radio link. The Base Station Subsystem communicates with the mobile service across the A interface.

2.4.1. Network architecture

2.4.1.1. Mobile station

The Mobile Station consisting of the physical terminal contains the radio transceiver, the display and digital signal processors and the Subscriber Identity Module (SIM). As seen in section 2.4.6.1, the SIM provides the user with the ability to access his subscribed services irrespective of the location and the terminal used. The insertion of the SIM in any GSM cellular phone allows the user to access a network, give and receive phone calls and make use of all the subscribed services.

The International Mobile Equipment Identity (IMEI) uniquely identifies the mobile terminal according to the International Mobile Subscriber Identity (IMSI) contained in the SIM. Because the IMEI and IMSI are independent, personal mobility is possible. The SIM can be protected against unauthorised use by a personal identity number (PIN).

2.4.1.2. *Base station subsystem*

The Base Station Subsystem is composed of two parts, the Base Transceiver Station (BTS) and the Base Station Controller (BSC). They communicate across the specified A-bis interface thus allowing network operators to use components made by different suppliers [M. Bezler et al, 1993].
The Base Transceiver Station houses the radio transceivers that define a cell and handle the radio link protocols with the Mobile Station. Depending on the density of the area, more or less BTS are needed to provide the appropriate capacity to the cell. DCS networks working at 1800 MHz need twice the number of BTS to cover the same area as GSM networks but provide twice as much capacity.

The Base Station Controller manages the radio resources for one or more Base Transceiver Stations. It handles radio channels set-up, frequency hopping, and handovers. The BSC is the connection between the mobile and the Mobile service Switching Centre (MSC). The BSC also takes care of converting the 13kbps voice channel used over the radio link (Um interface) to the standardised 64kbps channel used by the Public Switched Network (PSTN) or ISDN.

2.4.1.3. *Network subsystem*

The Mobile service Switching Centre (MSC) is the main component of the Network Subsystem. Its provides the same functionality of a switching node in a PSTN or ISDN but also takes care of all the functionality needed to handle a mobile subscriber such as registration, authentication, location updating, handovers and routing to a roaming subscriber. The MSC also acts as a gateway to the public fixed network (PSTN or ISDN) and signalling between the different entities uses the ITU-T signalling system number 7 (SS7), widely used in ISDN and current public networks.

The international roaming and call routing capability of the GSM networks are provided by the Home Location Register (HLR) and Visitor Location Register (VLR) together with the MSC. The HLR database contains all the administrative information about each registered user of a GSM network along with the current location of the Mobile Station. The current location of a Mobile Station is in the form of a Mobile Station Roaming Number (MSRN) and used to route a call to the MSC where the mobile is actually located. Even if only one HLR is found for each GSM network it is often implemented as a distributed database.

The Visitor Location Register contains a selection of the information from the HLR, basically all necessary information for call control and provision of the subscribed services, for each single mobile currently located in the geographical area controlled by the VLR. Even if each functional entity can be implemented as an independent unit, most manufacturers of switching equipment implement one VLR with its associated MSC, thus simplifying the signalling required.

Two other registers are used in GSM networks for authentication and security purposes. The Equipment Identity Register (EIR) is a database that contains a list of all valid equipment on the network, where each Mobile Station (MS) is identified by its International Mobile Equipment Identity (IMEI). An IMEI is marked as invalid when it has been reported stolen or the local GSM network does not approve its type. The Authentication Centre is a protected database that keeps a copy of the secret key stored in each subscriber's SIM card, which is used for authentication and ciphering of the radio channel.

2.4.2. The radio link

The International Telecommunication Union, allocated two frequency bands for the uplink (Mobile Station to base station) and for the downlink (base station to Mobile Station).for mobile cellular networks. The uplink uses the frequencies between 890Mhz and 915Mhz, when the downlink uses the 935-960Mhz band. Although this range has first been used by analogue networks the top 10Mhz of each band have been reserved for GSM networks by the CEPT. It is believed that eventually the entire 2x25Mhz bandwidth would be allocated for GSM (or its possible natural evolution).

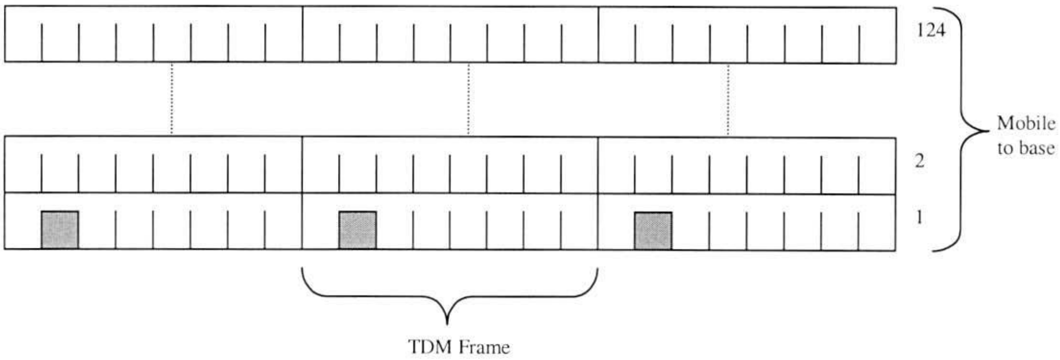


Figure 2: GSM channels and time slots

The radio spectrum being a limited and then precious resource, a method must be devised in order to make the most of the available bandwidth. GSM uses a combination of Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA). The FDMA part involves the division of the total 25Mhz bandwidth into 124 carrier frequencies of 200Khz each. Each base station is then assigned one or more of these carrier and each of these carriers are then divided in 8 time slots using a TDMA scheme. The Mobile Station uses 2 of these slot, one for emission and one for reception, separated in time so that the mobile does not have to receive and transmit at the same time.

2.4.2.1. *Channel structure*

The next figure represents the structure of one the time slot burst. Three other different type of time slot is used for frame and carrier synchronisation and frequency correction. A total of 156.25 bits are transmitted in 0.577 milliseconds, thus achieving a gross bit rate of 270.833 kbps. The 26 bits training sequence is used for equalisation as described in section 2.4.2.4, where as the 8.25 bit guard time allows for some delay in the arrival of the bursts. Each TDMA frame is constituted by a group of eight time slots, transmitted in 4.165 ms, groups of 26 or 51 TDMA frames also forms multi-frames to carry control signals. The 26 multi TDMA frame contains 24 traffic channels (TCH) and two Slow Associated Control Channels (SACCH) which supervise each call in progress. The SACCH are located in frame 12 and 25 of the 26 multi-TDMA frame. The SACCH contains eight channels, one for each of the connection carried by the TCHs. The SACCH in frame 25 is reserved for future use such as for half-rate traffic [C. B. Southcott et al, 1989].

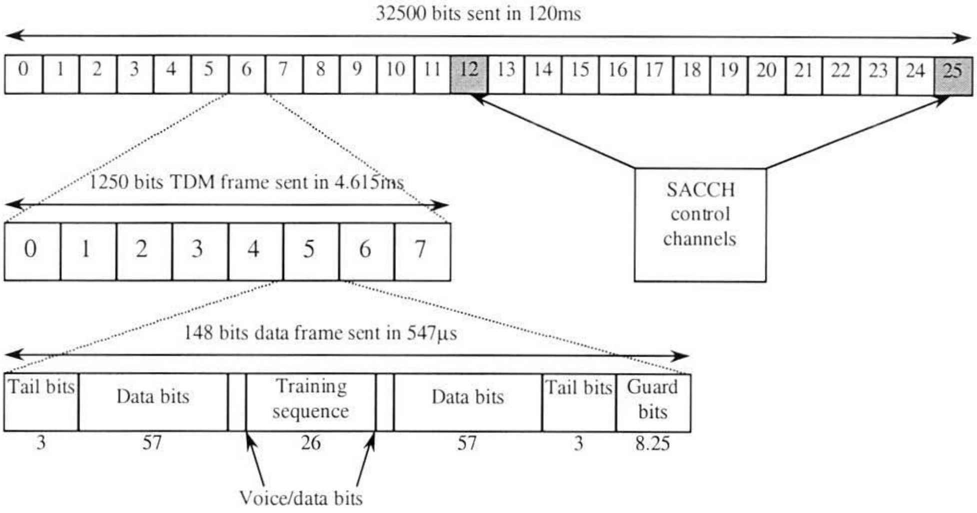


Figure 3: GSM framing structure

The 51 multi TDMA frame contains several control channels and is implemented on a non-hopping carrier frequency in each cell. These control channels includes:

- The Broadcast Control Channel (BCCH): used on the downlink to provide information about the base station identity, frequency allocations and frequency hopping sequences;
- The Stand-alone Dedicated Control Channel (SDCCH): is used for registration, authentication, call set-up and location updating;
- The Common Control Channel (CCCH): used during call origination and call paging;
- Random Access Channel (RACH): slotted aloha channel to request access to the network;
- Paging Channel (PCH): used to alert the mobile of an incoming call;
- Access Grant Channel (AGCH) used to allocate an SDCCH to a mobile for signalling.

2.4.2.2. *Speech coding*

Speech signals, inherently analogue have to be sample and converted to digital, in order to be carried onto the digital GSM network. Most current telephone systems and ISDN employ Pulse Coded Modulation (PCM) to multiplex voice lines over high-speed trunks and optical fibres. However the resulting bit rate of 64kbps is unsuitable for a radio link. The GSM group had to study several voice coding algorithm, and take into consideration subjective speech quality versus complexity, thus achieving a reasonable quality of service to cost ratio. They eventually choose Regular Pulse Excited - Linear Predictive Coder (RPE-LPC), which consider information from previous sample to predict the current sample. Speech is divided into 20ms samples each containing a total of 260 bits, the overall bit rate obtained is then 13kbps.

2.4.2.3. *Channel coding and modulation*

Each radio interface transmitting data is subject to natural man-made electromagnetic interference and thus needs to implement ways of protecting the transported information. The GSM system uses convolution encoding as well as block interleaving in order to achieve this protection. Different algorithms are used for voice or different data bit rates. This section will only describe the algorithm used for voice.

Experimentation has shown that some bits of the 260 bits of a voice sample are more sensitive to bit errors for perceived speech quality. Three classes have been defined to describe this:

- Class Ia 50 bits: most sensitive to bit errors.
- Class Ib 132 bits: moderately sensitive to bit errors.
- Class II 78 bits: least sensitive to bit errors.

Error detection for the Class Ia bits is performed using a 3 bits cyclic redundancy code. Should the 50 bits frame be judged too corrupted, it is discarded and replaced by a slightly altered version of the previous frame. These 53 bits together with the 132 Class Ib bits and a 4 bits trail sequence are encoded using an algorithm called half rate convolution encoder of constraint length 4. The algorithm works by encoding each input bit as 2 output bits, based on a combination of the previous 4 input bits. The encoder outputs 378 bits and the last 78 Class II bits are added to it without being protected. The overall process produces 456 bits for each 20ms sample. The GSM network further protects the sample by diagonally interleaving it. Eight blocks of 57 bits are created from the 456 bits sample output from the encoder and transmitted in eight consecutive time slot bursts. Each time slot burst can carry two 57 bits blocks and then two different speech samples are contained in each burst.

The GSM network uses Gaussian-filtered Minimum Shift Keying digital modulation to transmit the signal over the analogue carrier. GMSK has been chosen as a compromise between spectral efficiency and complexity of the transmitter. The less complex the transmitter is, the lower the power consumption is and then the greater the battery life.

2.4.2.4. *Radio signals reflection*

Radio waves at the 900Mhz (GSM) or 1800Mhz (DCS) frequency bounce off every natural or artificial object (cars, buildings, aeroplanes, etc.), thus many reflected signals differing by their phase reach the mobile station antenna. The technique used to extract the desired signal from all the reflections is called equalisation and works by finding out how a known signal is modified and constructing an inverse filter. The known signal is the 26 bits training sequence described in section 2.4.2.1 that is transmitted in the middle of every time slot burst.

2.4.3. *Power management aspects*

2.4.3.1. *Discontinuous transmission*

In a normal conversation it has been discovered that a person would speak less of 40% of the time. This consideration led to the use of Discontinuous Transmission (DTX), thus turning off the power of the transmitter during silence periods, and minimising the power loss at the mobile station side. Voice activity detection is the most important component of the DTX and must distinguish between voice and noise signals. The process is quite hard to implement as the misinterpretation of voice input as noise results in a very annoying effect called clipping [C. B. Southcott et al, 1989]. On the other end, if noise is misinterpreted as voice too often, the efficiency of the DTX is dramatically decreased. The digital nature of the GSM network leads to an uncomfortable silent at the receiving end when the transmitter is turned off. Comfort noise is created at the receiving end by analysing the characteristics of the transmitting end's background noise.

2.4.3.2. *Discontinuous reception*

The paging channel is used to alert the mobile station of an incoming call. It is structured in a way that the mobile station knows when to check for a paging signal, and can then go into sleep mode between two paging signals.

2.4.3.3. *Power control*

Different mobile stations are used on the GSM network, depending on their peak transmitter power. Five classes are defined and are rated from 20, 8, 5, 2, 1, and 0.8 watts. Both the mobile and the Base Transmitter Station operate at the lowest power level that would maintain an acceptable signal quality. The Base Station Controller decides when to change the power level according to the information that the mobile station passes to it.

2.4.4. *Network aspects*

The geographical area in the GSM networks is divided into cells, which requires the implementation of a handover mechanism that will be described in section 2.4.4.1. More over, the ability for a mobile station to roam nationally (between cells) and internationally (between networks and provided that the appropriate roaming agreements have been signed between the concerned networks), requires strong authentication, registration and call routing and location updating functions.

The GSM protocol implements the first 3 layers of the OSI model, the previous sections have described the channel structure forming the physical layer. Layer 2 is the data link layer and uses the LAPD protocol across the Um interface and Signalling System Number 7 across the A interface. The third layer is divided into 3 sub layers. The Radio Resource Management sub layer for set-up, maintenance and termination of radio channels, The Mobility Management sub layer for location updating, handover and registration and the Connection Management sub layer for general call control and supplemental services [MOBITEX].

2.4.4.1. Handover mechanism

The mechanism used to switch an ongoing all between different channels (more precisely time slots) or cells is called handover . Handover between cells can be done between cells controlled by the same Base Station Controller (BSC) or by different BSCs but controlled by the same Mobile Switching Centre (MSC) or even cells under the control of different MSCs. These four types of handovers are classified into internal (first two) or external handovers (last two) and do not involve the same kind of network resources. The internal handovers are handled by the concerned BSC which only notify the MSC at the completion of the handover. The external handovers are handled by the MSCs involved [D.M. Balston, 1991].

The MSC can initiate handovers as a mean of traffic load balancing or the Mobile Station can decide to start a handover. During its idle time slots, the Mobile Station scans the Broadcast Control channel of up to 16 of the neighbouring cells, and selects six of them for best candidates considering the signal strength. The Mobile Station then forwards the information to the BSC and the MSC to be used by the handover algorithm. Handover is generally considered when increasing the power level of the Mobile Station no longer increases the signal strength. This still is the simplest and most common method used and is called *"minimum acceptable performance"*. Other algorithms are being used but the decision whether a handover should be initiated is not part of the GSM specification.

2.4.4.2. Location updating and call routing

The Mobile Switching Centre (MSC) handles the interface between the GSM network and the public switched telephone network (PSTN). The PSTN network only sees the MSC as another one of its switching node, however from the MSC point of view the switching mechanism is much more complex, since the MSC has to know where the mobile is situated. The Mobile terminal could be anywhere in the home GSM network or even roaming on another GSM network in the world. Two specific registers are used in order for the MSC to accomplish location updating and call routing: the Home Location Register (HLR) and the Visitor Location Register (VLR) [S. Mohan et. Al. 1994].

The Mobile Station initiates a location updating when, by monitoring the Broadcast Control Channel, it notices that the location broadcast is not the same as the one previously stored in the mobile's memory. The Mobile Station then sends a request together with its International Mobile Subscriber Identity (IMSI) to update the new VLR via the new MSC. The VLR then allocates a Mobile Station Roaming Number (MSRN) and sends it to the Mobile Station HLR, where the most current location is stored. The HLR then sends back the appropriate call control parameters to the Mobile Station and informs the previous VLR that the MSRN can be reallocated.

The most common call routing procedure to a roaming Mobile Station can be described as follows: a call from the PSTN or ISDN network is placed to a mobile subscriber. According to the Mobile's Subscriber's telephone number (MISDN), the call is routed through the fixed land network to the appropriate GSM gateway. The gateway then sends a request to the HLR together with the MISDN. The HLR returns the current Mobile Station Roaming Number (MSRN). The MSC then routes the call to the MSC under which the destination Mobile Station is situated. The VLR of the current MSC finally converts the MSRN to the Mobile Station's IMSI and a paging call is broadcast by the cells under the control of the current BSC to inform the Mobile Station of the incoming call.

2.4.4.3. *Authentication and security*

Since the radio medium can be accessed by anyone, radio networks have to implement mechanism in order to protect the users. Authentication is one of the most important elements of the GSM networks [MOBITEX]. Authentication involves using two different entities, the SIM card in the Mobile Station and the Authentication Centre (AC). When subscribing to a GSM network the new user is given a secret key, one copy of which is stored in the Authentication centre. Authentication of a user on the network uses a ciphering algorithm called A3, the AC generates a random number and send it to the Mobile Station. Both the Mobile Station and the AC then use the random number together with the secret key and the A3 algorithm to generate a new number that the Mobile Station sends back to the AC. If the number generated by the Mobile Station and the AC are the same the user is authenticated. This calculated number is also used in conjunction with a TDMA frame number to encrypt the data sent over the radio link. The algorithm used in this case is called A5 and provides even greater protection to the already coded, interleaved and slotted data.

2.4.5. *Protocol architecture*

2.4.5.1. *GSM protocol layers*

The exchange of signalling messages regarding mobility, radio resource and connection management between the different entities of a GSM network is handled through the protocol architecture as shown on Figure 4. The architecture consists of three layers: the physical layer, the data link layer, and the message layer.

The message layer itself is divided in three sub layers on the MS side: the Connection Management (CM), Mobility Management (MM), and Resource Management (RM) sub layers.

- The CM sublayer consists of call-related supplementary services, Short Message Service (SMS), and call independent supplementary services support. SMS provides a connectionless data message service between the MS and a third party service centre;
- The MM sublayer provides functions to establish, maintain, and release a connection between the MS and the MSC over which an instance of the CM sublayer can exchange information with its peer. It also performs location updating, IMSI management, and Temporary Mobile Subscriber Identity (TMSI) identification, authentication and reallocation;

- The RR sublayer establishes the physical connection over the radio link to transmit call-related signalling information such as the establishment of the signalling and traffic channel between the MS and the BSS.

On the MSC side we still have the three layers, the higher layer being divided into four sublayers:

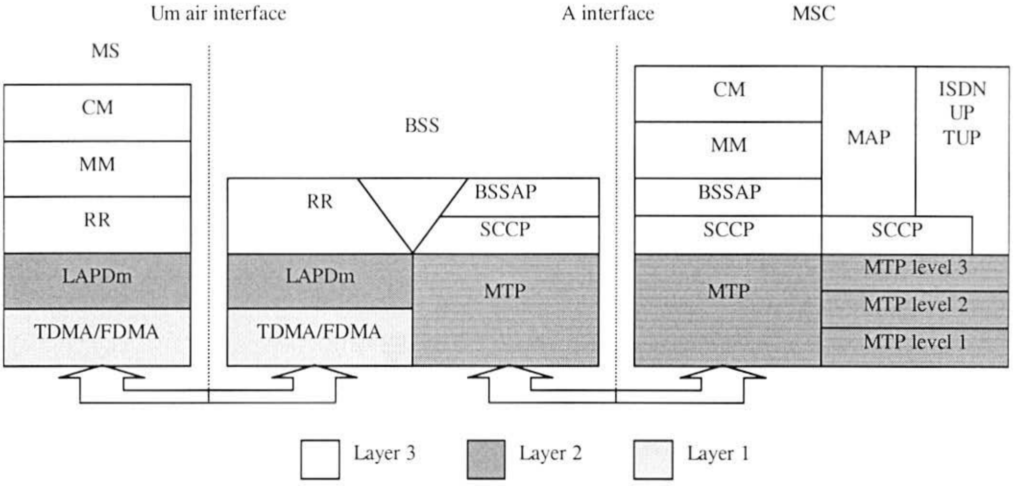


Figure 4: GSM protocol architecture

The Base System Substation Application Part (BSSAP) of the MSC provides the channel switching functions, radio resources management, and internetworking functions

The Message Transfer Part (MTP) and Signalling Connection Control Part (SCCP) protocols are used to implement the data link layer and the layer three transport functions for carrying the call control and mobility management signalling messages on the BSS-MSC link.

The SCCP packet is also used to carry the SMS messages. The LAPDm represents the data link layer over the radio link physical layer. It is based on the LAPD protocol that has been modified for operation within the constraints set by the radio path.

2.4.5.2. SS7 Protocol Layer

Signalling System Number 7 (SS7) is a protocol used by the telephone companies for interoffice signalling. It is currently the only element of GSM infrastructure capable of packet switching, as well as circuit switching. It is used to transport control signals and short message packet for the Short Message Service. The protocol consists of the Mobile Application Part (MAP), the Transaction Capability Application Part (TCAP), Signalling Connection Control Part, Message Transfer Part and the ISDN-User Part (ISUP) or Telephone User Part (TUP) [Jan A. Audestad, 1988].

The next figure shows a comparison between the OSI layer model and the SS7 protocol architecture [Rahnema 1993].

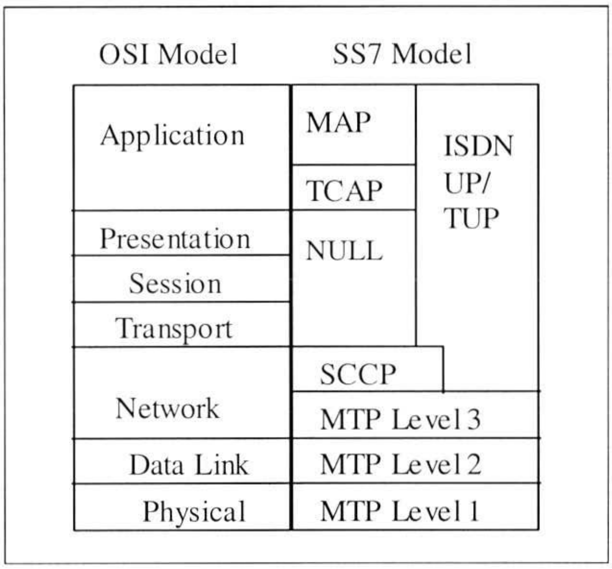


Figure 5: SS7 Protocol Architecture

The ISDN User Part (ISUP) provides the signalling functions needed to support switched voice and data application in the ISDN environment.

The Telephone User Part (TUP) provides the basic functionality for call control functions for ordinary national and international telephone call.

The Transaction Capability Application Part (TCAP) is an application layer protocol. It allows an application at one node to invoke an execution of a procedure at another node and exchange the results of such invocation. It isolates the user application from the complexity of the transaction layer by automatically handling transaction and invocation states changes, and generating the abort or reject messages in full accordance with ITU/CCITT and ANSI standards. The MAP uses the TCAP services to provide the signaling capabilities required to supporting the mobile capabilities [Mouly *at al.* 1995].

The MTP and SCCP correspond to the lower three layers of the OSI model. The SCCP sublayer, as shown in Figure 6, supports connectionless and connection-oriented services to transfer data. The data transfer is reliable, independent of the underlying hardware, and transparent to users. The protocol employs logical signalling connections within the SS7 network to ensure reliability and integrity of the ongoing data transfer. It also supports global title translation for voice, data, ISDN, and GSM services. The MTP is divided into three layers:

- MTP level 1 transfers signalling messages over the network and acts as a signalling transfer point;
- MTP level 2 is equivalent to the OSI data link layer and provides a reliable sequenced delivery of data packets across MTP level 1;

- MTP level 3 provide congestion control, signalling management, and message discrimination, distribution and routing.

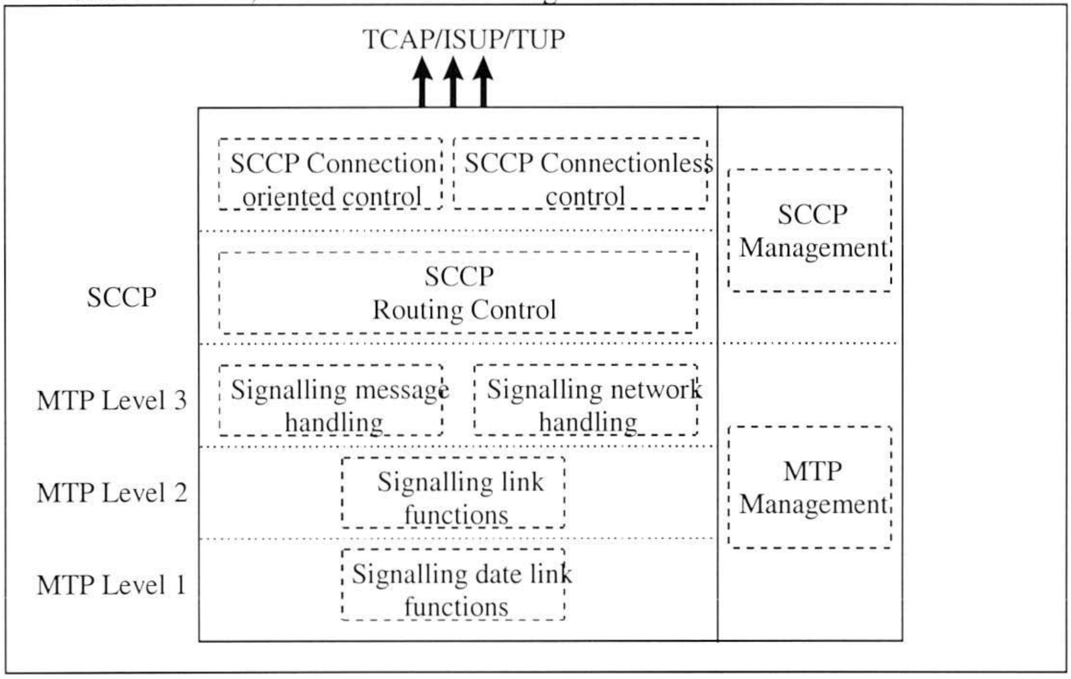

TCAP/ISUP/TUP
SCCP Connection oriented control
SCCP Connectionless control
SCCP Management
SCCP
SCCP Routing Control
MTP Level 3
Signalling message handling
Signalling network handling
MTP Level 2
Signalling link functions
MTP Management
MTP Level 1
Signalling date link functions

Figure 6: The structure of the SS7 SCCP and MTP layers

2.4.6. *Features provided*

2.4.6.1. *Security*

GSM provides powerful signalling capabilities that facilitate and enhance international roaming. When travelling abroad and provided that a roaming agreement between the home and roaming network has been signed, the user has access to the roaming network through automatic network location detection and registration.

Personal mobility is achieved through the insertion of a subscriber identity module (SIM) into the GSM network. The SIM carries the personal number and subscription details assigned to the cellular network user as well as storage space for a calling directory. User authentication is done by the home network database whether the user is on his home network or roaming abroad.

Using digital technology provides GSM networks users with exceptional security. The use of full digital encryption greatly enhances security and applies to voice and data calls, thus preventing false charge on the user's bill, ensuring that incoming calls are delivered to the right phone and reducing the risk of eavesdropping drastically. Digital phones are inherently difficult to eavesdrop for the multiple encryption mechanisms used are virtually unbreakable without the appropriate equipment.

2.4.6.2. *Performance*

Whilst many of the current analogue systems can offer good performance, the GSM system has been designed to outrank analogue systems. Speech quality on GSM networks is comparable under average to good conditions; however GSM performs significantly better under weak signal or bad interference.

Due to the digital standards employed, a high level of integration has been achieved leading to smaller; lighter handsets and using low-voltage technologies manufacturers are improving battery life significantly. Nowadays credit card size GSM mobile phones with a standby battery life of over 2 weeks are starting to be widely available.

GSM greatly improves spectrum efficiency compared to analogue networks and thus provide greater network capacity. The performance of a cellular network is mainly restricted by co-channel interference and this is greatly reduced by using digital technology.

2.4.6.3. Mobile data services

Communication Standards have always been classified into two categories: packet switched or circuit switched protocols. In cellular networks, circuit switched protocols are mainly used: data flows through an end-to-end analogue channel as compatible modems at both ends ensure the modulation and demodulation. Special protocols (e.g. V.22bis, V.34) have been developed in order to help the modem maintaining the cellular connection for the duration of the data transfer. The cellular modem using these protocols provides error correction and can handle cell hand-offs but does not account for congestion control or packet routing. Packet switching belongs to the digital world and uses well-known protocols such as X25 or TCP/IP, but has proved to be less than ideal over a wireless link (even after modifications and optimisation). **[Mobeen *et al.* 1995].** Wide area wireless packet network manufacturers are now taking a brand new approach and are working on developing new protocols, specifically designed to tackle the unreliability of the radio environment while still providing the reliability and efficiency of packet-switched networks. Ericsson's MOBITEX, Motorola's RD-LAP, the Cellular Digital Packet Data (CDPD), High Speed Circuit Switched Data (HSCSD) and General Packet Radio Service (GPRS) are examples of this strategy and will be introduced shortly in this chapter.

2.4.6.3.1. Ericsson's MOBITEX

Developed by Ericsson more than 10 years ago, MOBITEX is a mobile data technology that has been improved over the years. Originally mixing low-bandwidth data transfer with voice, the protocol has been enhanced in 1990 for use in North America and the United Kingdom and is now a data only system. The enhanced version uses the cellular architecture and multi-channel frequency reuse as well as GMSK modulation, Hamming code for error correction and a reservation slotted ALOHA type scheme to access the channels. It is now the most widely deployed packet switched data technology (14 operators in 13 countries).

2.4.6.3.2. Motorola's RD-LAP (ARDIS)

Originally created as a private network for IBM the system is now available for public use [RD-LAP]. It uses a concept known as single frequency reuse where all cells in a given area share the same single frequency. This allows the system to be deployed over very little spectrum but compromises capacity. ARDIS was originally deployed using MDI's (Mobile Data international Inc.) MDC-4800 protocol (4.8Kb/s in a 25 KHz channel) before Motorola developed the RD-LAP protocol that offers 19.2Kb/s in the same channel. Actually available in 7 countries the deployment of the ARDIS network is still very limited.

2.4.6.3.3. Cellular Digital Packet Data

Introduced by IBM as a packet switching overlay for the existing analogue cellular voice network [CDPD]. The original system was intended to make use of idle voice channels to multiplex short data messages and to hop among frequencies to find these idle channels. However, the system quickly showed its limitations: overall complexity of the process and the interference caused reducing the voice quality. This caused a revision of the specification which now supports IP and OSI protocols at the network layer and adds its own robust physical and link layers for the air interface, thus making it easier to port existing applications for use in the mobile environment.

2.4.6.3.4. HSCSD

High Speed Circuit Switched Data [HSCSD] is an enhancement of data services ("Circuit Switched Data - CSD) of all current GSM networks. It allows you to access data services at 3 times faster, which means subscribers are able to send and receive data from their portable computers at a speed of up to 28.8 kbps; this is currently being upgraded in many networks to rates of and up to 43.2 kbps. The HSCSD solution enables higher rates by using multiple channels, allowing subscribers to enjoy faster rates for their Internet, e-mail, calendar and file transfer services.

2.4.6.3.5. GPRS

The General Packet Radio Service [GPRS] is another data value added service that allows information to be sent and received across a mobile telephone network. It supplements today's Circuit Switched Data and Short Message Service.

2.4.6.3.6. WAP

The Wireless Application Protocol [WAP] is a hot topic that has been widely hyped in the mobile industry and outside of it. WAP is simply a protocol- a standardised way that a mobile phone talks to a server installed in the mobile phone network. The Wireless Application Protocol (WAP) is an important development in the wireless industry because of its attempt to develop an open standard for wireless protocols, independent of vendor and air link. It has however received mixed success, partly due to the lack of handsets at first and then the relatively slow connections achievable using today's GSM data services. GPRS should hopefully turn it around.

2.4.6.4. The short message service

Developed as part of the GSM Phase 2 specification, the Short Message Service, or SMS as it is more commonly known, is based on the capability of a digital cellular terminal to send and/or receive alphanumeric messages. The Short Messages can be up to 160 characters in length, and are delivered almost instantly and anywhere in the world (where GSM coverage is available). What's more, the delivery of the message is guaranteed even when the cellular terminal is unavailable (e.g. when the terminal is switched off or outside the area of coverage), i.e. the network will hold the message and deliver it shortly after the cellular terminal announces its presence on the network.

There are several ways in which an SMS could be sent or received, for example, using a central paging service (i.e. an operator), or directly from an SMS enabled terminal (Mobile originated). Many networks also currently support modem dial-up access to the service allowing PC users to send and/or receive messages using either a modem (PSTN) or a GSM data card. The fact that SMS (through GSM) supports international roaming with very low latency makes it particularly suitable for applications such as paging, voice mail notification, messaging services for multiple users, etc. However, the facilities offered to users and the charges for these facilities still mainly depend on the level of service provided by the network operator.

There are two types of short message service available: Cell broadcast and Point-to-Point. In cell broadcast, a message is transmitted to all the active MS present in a cell that have the capability of receiving SMS and have subscribed to that particular information service. This service is only one way and no confirmation of receipt will be sent. It can send up to 93 7-bit character or 82 8bits characters **[ETSI 3.41 1996]**. It is used to transmit messages about traffic conditions, weather forecast, stock market, and other useful information to the MS.

In Point-to-Point service, messages can be sent from one mobile to another or from a PC to a mobile and vice versa. These messages are maintained and transmitted by an SMSC. The SMSC is an electronic form of ordinary mail postal service that will store-and-forward the messages until they can be delivered. Each GSM network will provide one or more SMSC to sort and route the messages. Each SMSC will check, organise, and send the message to the operator. It will also receive and pass on any confirmation message to any GSM mobile on any network.

2.5. Conclusion

The GSM standard is a complex mixture of technology and features unprecedented in the mobile communication technology. It provides high quality voice conversation to the end user together with a high level of security and the ability to seamlessly handover between cells and roam between networks. It has seen an exponential growth in terms of both user base and number of networks in operation throughout the world since it was first launched with now over 400 networks available and millions of subscribers. The Short Message Service has probably seen the most amazing growth any text based information and communication service has seen with over 50 billion messages exchanged in the UK

itself in the first quarter of 2001. From a simple voicemail notification service, the SMS is now widely used for a variety of purposes including the delivery of binary multimedia content such as ring tones, logos or WAP service settings configuration.
The next chapter focuses on the architecture of the SMS and introduces a few of the many protocols used to interface with the Short Message Service Centres (SMSC).

3. The short message service

3.1. Introduction

Developed as part of the GSM Phase 2 specification, the Short Message Service, or SMS as it is more commonly known, is based on the capability of a digital cellular terminal to send and/or receive alphanumeric messages. The short messages can be up to 140 bytes in length, and are delivered within a few seconds where GSM coverage is available. More than a common paging service, the delivery of the message is guaranteed even when the cellular terminal is unavailable (e.g. when the terminal is switched off or outside the area of coverage). The network will hold the message and deliver it shortly after the cellular terminal announces its presence on the network.

The fact that SMS (through GSM) supports international roaming with very low latency makes it particularly suitable for applications such as paging, e-mail or voice mail notification, messaging services for multiple users, etc. However, the facilities offered to users and the charges for these facilities still mainly depend on the level of service provided by the network operator.

There are two types of short message service available: Cell broadcast **[ETSI 3.41 1996]** and Point-to-Point **[ETSI 3.40 1996]**.

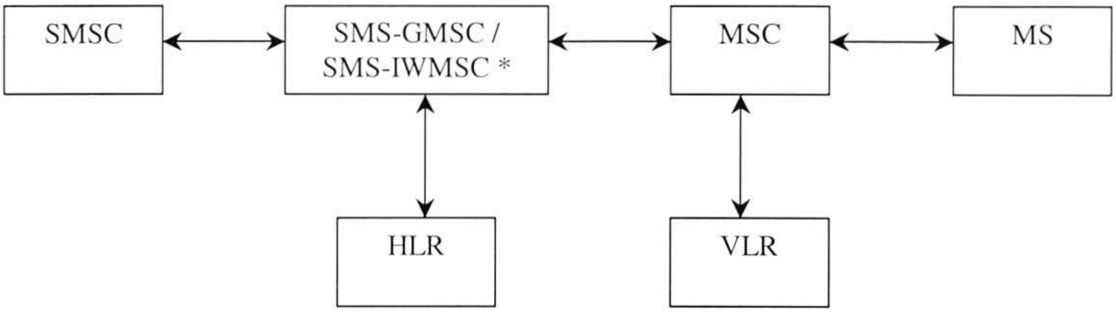
3.2. Practical implementation

The Short Message Service uses the signalling channel SS7 to transmit the data packets **[ROTH 1993]**. Thus, allowing a text message to be received when the user is making a voice or data call. An active Mobile Station (MS) should be able to send and receive a short message Transport Protocol Data Unit (TPDU) at any time regardless of whether there is a speech or data call in progress. A confirmation will always be returned to the Short Message Service Centre (SMSC) indicating whether the MS has received the short message or not. A confirmation will also be returned to the MS from an SMSC indicating whether the TPDU has been received successfully. The software within the MS must be able to decode and store the messages.

SMS Mobile Terminated (SMS-MT) is the ability to receive a SMS message from a SMSC and is more ubiquitous, while SMS Mobile Originated (SMS-MO) is the ability to send a SMS message to SMSC. Messages can also be stored on the SIM, which can be retrieved at a later time. When the phone is not on service area or the SIM is full, the SMSC will hold the message and deliver it shortly after the phone comes back into range or there is a space in the Memory. A SMS message can also be sent using the public dial-up access provided that the network operator offer it. Three of the four UK networks have public dial-up access, either using plain text or Telocator Alphanumeric Protocol (TAP).

3.2.1. Basic network architecture

The SMS protocol layer architecture is depicted on the figure below.



*: SMS-GMSC for mobile terminated SMSC to MS)
SMS-IWMSC for mobile originated (MS to SMSC)
They can be integrated with the SMSC

Figure 7: Basic network structure

When routing a mobile originated short message, the SMSC forwards the short message to the SMS-GMSC. The SMS-GMSC interrogates the HLR for routing information and sends the short message to the appropriate MSC. The MSC delivers the short message to the MS.

On the other hand, when routing a mobile terminated short message the MS addresses the required SMSC as specified by an E.164 address (message centre international format number). If roaming abroad the visited PLMN will route the short message to the appropriate SMS-IWMSC.

The Short Message Service Centre (SMSC) identifies each short message uniquely by adding a time stamp in the SMS-DELIVER TP-SCTS field. The short message arrival at the SMSC is accurate to the second. Consequently, it is the SMSC responsibility to assure that if two or more short message arrive within the same second their time-stamp will be different.

The MS has to be able to receive/submit a short message TPDU that handles the delivery report or deliver one upon successful reception. It is also responsible for notifying the network when it has memory capacity available to receive one or more messages, if it had previously rejected a short message because its memory capacity was exceeded.

3.2.2. *Protocol architecture*

The protocol layer for the short message service is depicted on the picture below:

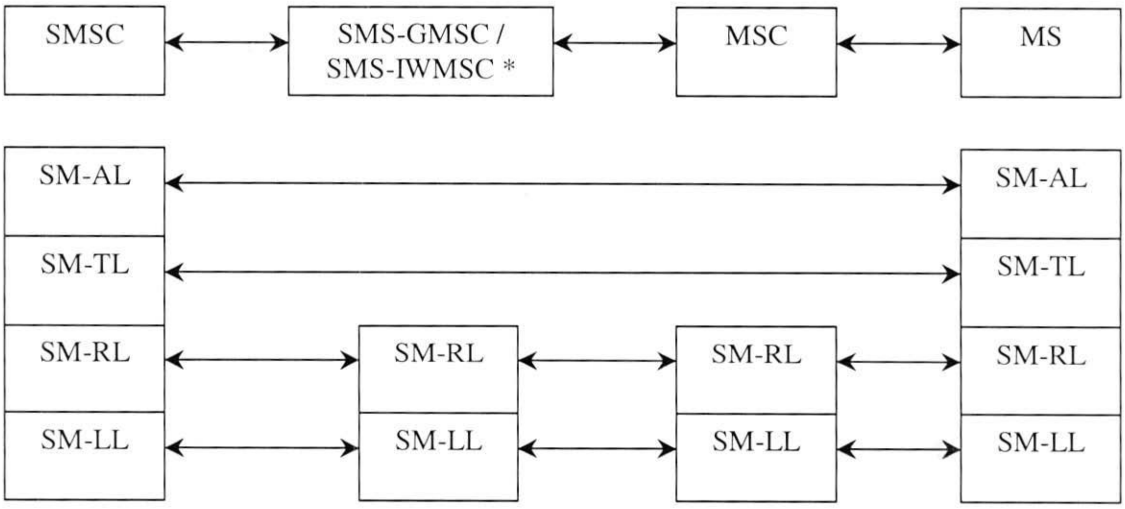


Figure 8: Protocol layer overview for SMS point to point

3.2.3. *SMS protocol data unit types*

There are six types of Transfer Protocol Data Unit (TPDU):

- SMS-Deliver: conveying a short message from the SMSC to the MS;
- SMS-Deliver-Report: conveying a failure cause (if applicable);
- SMS-Submit: conveying a short message from the MS to the SMSC;
- SMS-Submit-Report: conveying a failure cause (if applicable);
- SMS-Status-Report: conveying a status report from the SMSC to the MS;
- SMS-Command: conveying a command from the MS to the SMSC.

The actual format of the SMS-Deliver and SMS-Submit TPDU are shown in the figure below **[ETSI 03.40 1996]**. These TPDU are used to carry the user's data and are the only that would be discussed in this document. The most important fields related to the submission and reception of short messages are described in the following sections.

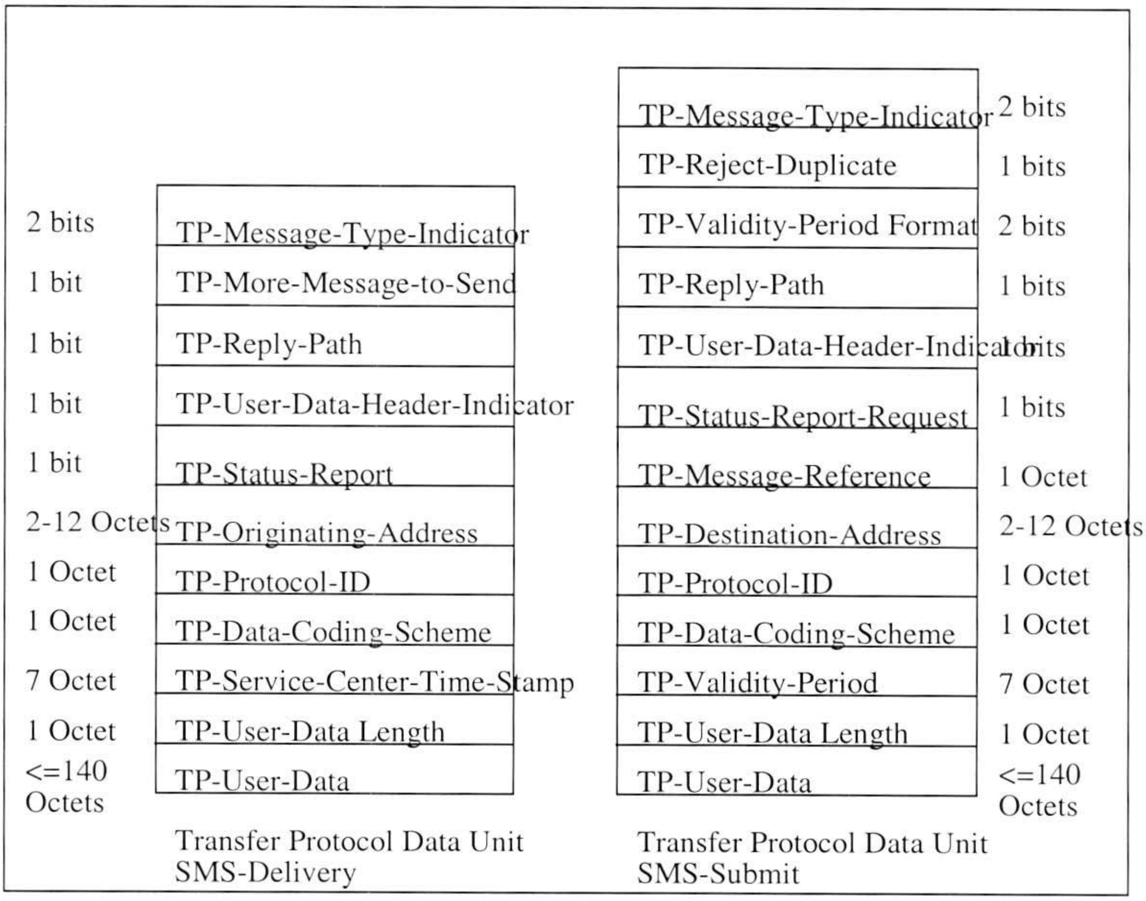


Figure 9: SMS Transfer Layer Deliver and Submit Protocol Data Unit

3.2.3.1. *TP-Protocol Identifier (TP-PID)*

This identifier is used by the MS or the SMSC to identify the higher layer protocol being used for internetworking with a certain type of telematic device (Telefax group 3 or 4, Ermes, etc.)

3.2.3.2. *TP-Data-Coding-Scheme (TP-DCS)*

The data coding scheme field is used to identify the coding scheme used by the user data, which can be 7bit or 8 bit or even Unicode (see section 4.3 for further details).

3.2.3.3. *TP-Validity-Period (TP-VP)*

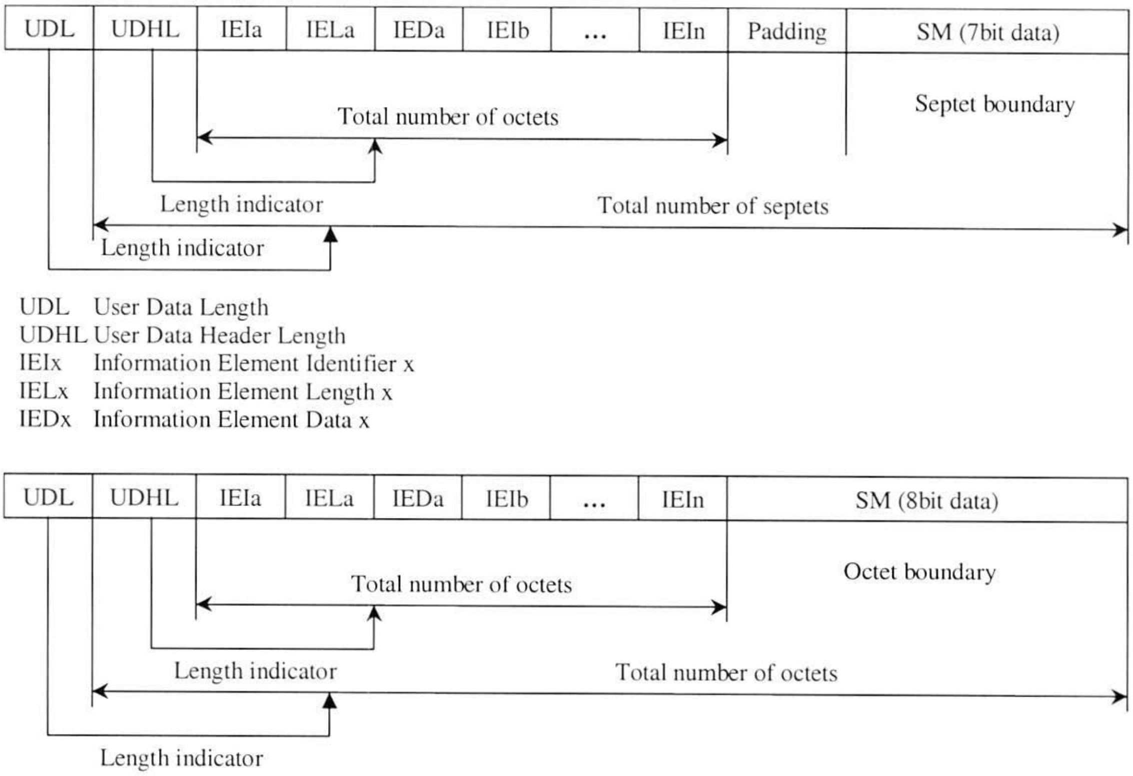
This field contains an information element enabling a MS to specify a validity period for the short message it is submitting. The value specifies how long an SMSC shall guarantee the existence of a short message before delivery to the recipient has been carried out.

3.2.3.4. *TP-More-Message-To-Send (TPMMS)*

The SMSC uses this field to inform the mobile station that one or more short messages are waiting to be delivered.

3.2.3.5. *TP-User-Data (TP-UD)*

The TP-UD field is used to carry the actual short message. It can store up to 140 octets of data for point to point SMS, together with a header depending on the setting of the TP-User-Data-Header-Indicator (TP-UDHI). The amount of space taken by the header reduces the number of characters the PDU can carry. The diagram below shows a representation of the layout of the TP-UD for 7 and 8bit data scheme.



UDL	User Data Length
UDHL	User Data Header Length
IEIx	Information Element Identifier x
IELx	Information Element Length x
IEDx	Information Element Data x

Figure 10: TP-USER Data Field Layout for 7 and 8 bit data coding scheme

The header will have at least 3 fields. The first field, the Information Element Identifier, is used to identify concatenated short messages. The Information Data Length is used to indicate the length of the Information Element Data that follows. Each of these fields is 1 octet long.

In the user data, the message can be 7 bit, 8 bit or 16 bit. If 7 bit data is used and the header does not end on a 7-bit boundary then padding bits shall be used. This is to ensure that older mobiles that do not support the TP-UD header can still display the message properly. The maximum length of the short message itself varies according to the data coding scheme used and the presence of one or more headers (134 and 152 octets for 8 and 7 bits respectively).

3.2.4. *Concatenated Short messages*

Using the Information Element Identifier, concatenated short messages can be sent and received. The maximum length of the message is then increased to 38760 (255*152) or

34170 (255*134) depending on the coding scheme used. The Information Element Data field contains all the necessary information for the receiving entity to re-assemble the messages in the correct order, and shall be coded as specified below.

- First octet: short message reference number identifying the message within the same transaction;
- Second octet: specifies the maximum number of short messages in the concatenated short message. The value specified here will not exceed 255;
- Third octet: identifying the sequence number of the short message within the concatenated message.

3.2.5. Binary content

Binary data carried by the SMS is a fairly recent use the of service. In this case, non textual information is carried in the 140 bytes available and can span multiple messages to carry the whole payload. Headers are included in the payload of the messages to carry information related to the re-assembly of the data by the receiving end. Using the SMS as a bearer for this type of data is an interesting and efficient way of transporting and remotely configuring handsets for mobile data services (such as WAP or GPRS settings) or simply deliver multimedia items such as ring tones or logos.

3.3. Short message routing

Figure 11 shows how a single SMS message is transmitted. A singlee arrow line indicates the SMS message, while double dotted arrow line indicates the confirmation message. A double arrow line indicates normal conversation. Mobile A is sending an SMS message to C while a voice call is in progress between B and C. The SMS message will pass through a series of SCCP switches, which can be within the MSC before reaching the SMSC and Visiting MSC.

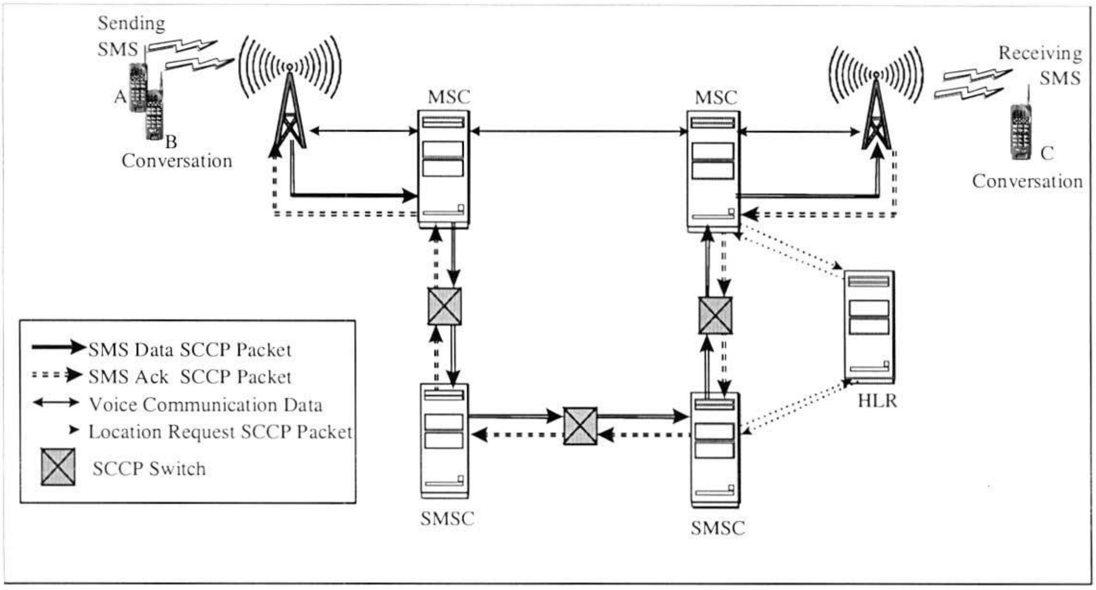


Figure 11: Basic short message routing

When a mobile originated SMS message is sent, the local cellular exchange or the Visiting MSC (VMSC) will route the SMS message in a Signalling Connection Control Part (SCCP) packet using the SMSC address. This address is normally stored on the phone or on SIM card. The international SCCP messages are routed based on a Global Title: the SMSC address. This SMSC address is defined by the E.164 numbering plan **[CCITT E.164 1997]**. The SCCP routing will be set-up in the VMSC, pointing to the local SCCP transit switch. Then the SCCP packet is passed from one switch to another until it reaches the destination SMSC. Each exchange will inspect the Global Title to make sure it is valid and use this to route the message to the next exchange in the chain. Once the message arrives at the destination SMSC, it will send a confirmation message back to the handset using another SCCP packet.

When the SMSC wants to send that message to a receiving mobile, it must find the location of the mobile. It will need to request the mobile's HLR database for location information. A location request SCCP packet, based on the mobile number, will be sent by the SMSC. This international SCCP network will then route the location request SCCP packet to the appropriate HLR. SCCP routing information must be in place in the international SCCP transit switches for all codes allocated to subscribers as well as the ones allocated to switches. When the HLR receives the request, it will return the location information in another SCCP packet to the SMSC. The SMSC then sends the message to the VMSC of the mobile, based on the information received from the HLR. This VMSC will eventually send the message to the mobile phone using the cellular exchange global title. Once the mobile has received the message, the VMSC will send a confirmation SCCP packet back to the SMSC.

Throughout these routing procedures, message and acknowledgement can get lost if one of the cellular exchanges along the route does not know where to pass the SCCP packet to. A message can also be lost if the SMSC did not understand the message format. If the message did not reach the SMSC, the message will be sent again by the sender. If a message is sent twice, the repeated TPDU will be removed by the SMSC. Sometimes the message reaches the receiver, but the acknowledgement is lost. This can either cause the same message to be delivered to the handset multiple times or a negative acknowledgement to be received by the sender. Incorrect routing usually results in packet loss.

The complexity of routing messages increases widely when combined with roaming as explained in the following example: In Figure 12, User A in Network ① is sending a short message to User B in Network ③ roaming in network ④. User A is using the SMSC in Network ② to submit his short message.

The local cellular exchange routes the short message in an SCCP packet according to the SMSC Global Title as defined by the E.164 numbering plan [CCITT E.164 1997]. The SCCP packet is forwarded from exchange to exchange until it reaches the destination SMSC. The routing has to be set up in all the SCCP switches along the route for the message to successfully reach the SMSC in Network ②.

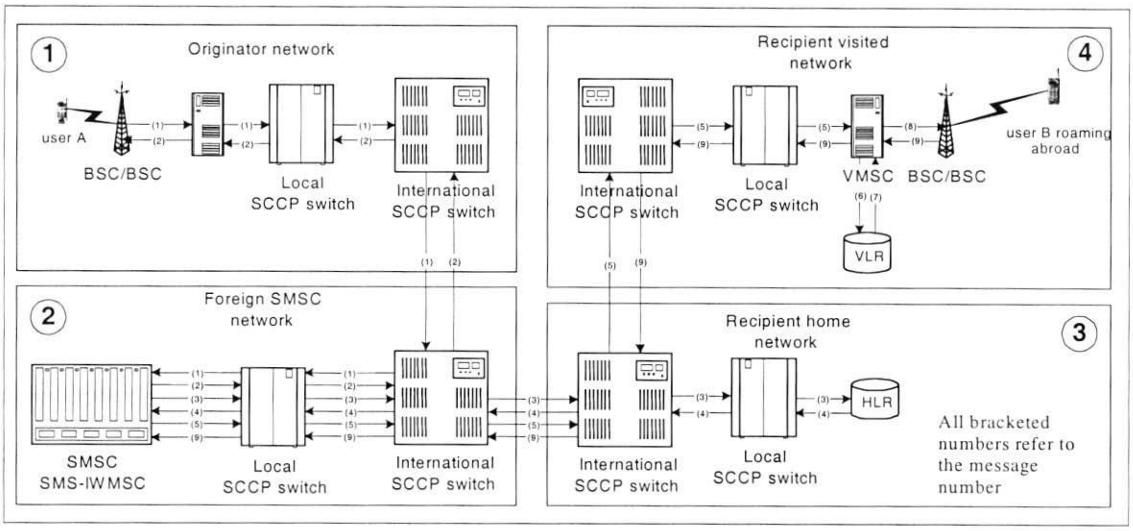


Figure 12: Complex short message routing scenario

Once the SCCP packet carrying the message arrives at the destination SMSC, a confirmation message is send back to the handset using another SCCP packet (2).

To deliver the short message to User B the SMSC has to access the HLR database of his home network. A location request SCCP packet, based on the User B mobile number is sent by the SMSC (3).

This international SCCP network then routes the location request SCCP packet to the appropriate HLR. When the HLR receives the request, it will return the location information in another SCCP packet to the SMSC (4).

The SMSC then sends the message to the VMSC of the User B, based on the information received from the HLR (5). Finally this VMSC interrogates the VLR (6) & (7) and delivers the message to User B (8). Upon successful delivery a confirmation SCCP packet is sent back to the SMSC (9).

Throughout these routing procedures, the SCCP packets can get lost if one of the cellular exchanges along the route does not know where to forward the SCCP packet. SCCP routing is based on the Global Title used for switches and the SMSC. The routing information has to be in place in the international SCCP transit switches for the messages to successfully reach their destination. Some international switches only check the country code prefix (e.g. 44 for the UK) and forward the packet to the next exchange, while others also check for the network prefix (e.g. 44976 for Orange). If the exchange routing table does not include all the prefixes allocated to the subscribers some messages will be rejected. Incompatible implementation of the SMSC can also lead to the short message not being understood and being rejected. All the above mentioned problems can lead to packets getting lost along the way with different consequences:

- Negative acknowledgement received by the sender (phone displaying "message failed" or a similar message) although the short message reaches its destination (loss of packet ②);

- Reception of duplicate short messages by User B (loss of packet ⑤ or ⑨), could also be due to the time out value being set too low in the SMSC;
- Or in the worst case the message might not be delivered at all (loss of packet ①, ③, ④, ⑤).

Even if Figure 12 shows one of the most complicated routing scenario, there is still much that can go wrong with short message.

3.4. Protocols for short message submission

The ETSI specified a protocol for short message submission as part of the overall GSM standard **[ETSI 7.05 1996]**. This specification defines three interface protocols for the transfer of SMS Short Messages between a mobile station (MS) and Terminal Equipment (TE) via an asynchronous interface. The protocols clearly overlap in functions and it is not clear why three have been defined.

3.4.1. Block Mode

The Block mode is a binary protocol which encapsulates the SMS PDU used for short message transfer between a mobile station and the SMSC which are defined in GSM 03.40 [ETSI 3.40 1996]. This protocol includes error detection and is suitable for use where the link between the application and the phone is subject to errors. It will be of particular use where control of remote devices is required. The application has to construct a binary string including a header and the short message PDU (SMS-TPDU).
Once the application has requested the phone to enter block mode a group of functions is available:

- Submit a short message;
- Delete messages from the phone;
- List messages in the phone;
- Transfer all or one message from the phone to the application;
- Set the phone so that the application is notified every time a new short message is received.

Each of these commands contains a number of predefined elements as described in the specification. For example, the *Insert Message* command format used to submit a short message is depicted in Table 1.

Table 1: Short message submission using Block mode

Information element	Meaning	Length Byte
Message Type	Insert SMS type: the value defined in the specification is 0x07 in hexadecimal or 00000111 in binary	1
Insert Type	1. Store in phone 2. Send, or 3. Store and send	1
RP-Dest Address	Address of recipient as defined by GSM 04.01	1-12
SMS-TPDU	As defined by GSM 03.40	Max 164

3.4.2. *Text mode*

The Text mode is a character-based protocol based on the “AT” commands set modified for GSM. This mode is suitable for unintelligent terminals or terminal emulators, and for application software built on command structures like those defined in ITU V25ter. The application passes the message in plain text to the phone that constructs the TPDU (see Table 2). This however means that the Text mode offers a lot less functionality than the Block mode or the PDU mode. The Text mode does not support, neither can it automatically pass incoming messages to the application (only notify it).

Table 2: Short message submission using Text mode

AT+CMGS=”44976123456”<CR> This is a text message<CR> ^Z	Send a message to 44976123456
+CMGS=3 OK	Message accepted by the phone with a reference number 3

3.4.3. *PDU mode*

The PDU mode is very similar to the Text mode, except that it leaves to the application the responsibility to build the short message TPDU. This mode adds to the convenience of the AT command set the possibility to construct more sophisticated PDUs (i.e. allowing binary data to be transmitted, not just characters).

3.4.4. *SEMA SMS2000*

Sema Group Telecoms developed SMS2000 as an implementation of a GSM SMSC. The specification mainly describes the delivery of short messages to mobile stations (MS), but also specifies the protocols for short message submission. The protocol has been designed to operate over a variety of interfaces such as X25, DECnet or SS7. The SMS2000 SMSC is usually accessed via the general X25 access gateway –either using a radio Packet Assembler Disassembler (PAD) or a dedicated link to the message centre.

Once connected to the SMSC a Subscriber Mobile Entity (SME) can request any of the following operations:

Table 3: SMS2000 commands

Submit SM	Send a SM to a MS
Delete SM	**Delete a previously submitted SM**
Replace SM	**Replace a previously submitted SM to a MS.**
Delete all SM	**Delete all previously submitted and undelivered SM to a MS**
Enquire SM	**Request status of a previously submitted SM**
Cancel SRR	**Cancel all Status Reports requests (SRR) about a previously submitted SM.**
Alert SME request	**Request to be alerted when a specified SME becomes registered.**
Retrieve Request	**Request transmission from the SMS2000 SMSC of any pending SM or SR.**
Login	**For X25 general access when accessing from a different location**
Change Password	**For X25 general access when accessing from a different location**

The SMS2000 SMSC can also send the commands listed on Table 3 to an SME. A transaction between the SME and the SMSC involves one of the party to send a request with a status report being sent back on completion or failure of the request.

Table 4 depicts the submission of a short message from an SME to the SMS2000 SMSC.

Table 4: additional SMS2000 commands

Alert SME	Indicates a MS has registered with the GSM network
Status Report	Indicates successful delivery or failure of a previously submitted SM.
Incoming SM	Indicates an incoming SM is being held by the SMS2000 SMSC

The transaction is initiated by the SME, a Submit SM invoke is sent to the SMSC. The SMSC responds with a result message indicating that the short message has been accepted and is being processed. Upon delivery the SMSC notifies the SME (if a status report has been requested). The SME then acknowledges the SR thus completing the transaction.

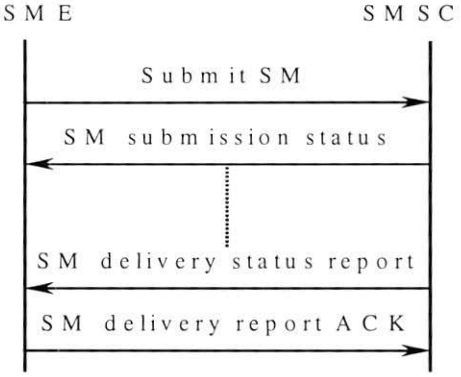


Figure 13: Short message submission using SMS2000

The SMEs connected to the SMS2000 SMSC are assumed to be trusted systems, a basic transaction will not include any exchange of login and password between the SME and the SMSC. However a login facility is still provided in order to access the SMSC from a different location (i.e. PAD).

3.4.5. Text based protocols

Usually these protocols are proprietary and developed as an interface to the SMSC of a digital cellular network operator. The advantage of a text based protocol is that the user does not need any special client software to submit a short message, they can dial-up to the appropriate message centre using any terminal emulation software and submit short messages using the different options offered. Figure 14 describes the submission of a short message using the Telenote text based protocol.

There are however key disadvantages with text based protocols: they offer limited support for extended character sets, and only work one way, to name a few. The user is only able to send messages and receive confirmation of submission. The SMSC is unable to notify the end user of successful delivery.

```
PLEASE  ENTER  THE  MOBILE  NUMBER  IN
INTERNATIONAL       FORMAT,        E.G.
44385XXXXXX,FOLLOWED  BY  RETURN,  (OR
RETURN TO QUIT).

>44374164828

PLEASE  TYPE  YOUR  TELENOTE,  (MAXIMUM
160 CHARACTERS), FOLLOWED BY RETURN.

>This is a test message

MESSAGE ACCEPTED
```

Figure 14: Short message submission using Telenote

3.4.5.1. Telocator Alphanumeric Protocol

Developed by Telecom Securicor Cellular Radio Limited, the Telocator Alphanumeric Protocol **[TSCR 1995]** provides greater flexibility and more features than the text based protocols. The overall performance is also significantly more efficient.
In its fully featured implementation, the protocol allows the user to perform the following operations:

1. Submit a short message and receive confirmation of acceptance;
2. Submit a short message and receive status of 1st delivery attempt;
3. Query the current status of a message submitted by (1) or (2);
4. Delete a message submitted by (1) or (2);
5. Replace a message submitted by (1) or (2), unless the message has already been delivered to the mobile;
6. Update a message submitted by (1) or (2) – If the message is still in the SMSC then it is replaced, otherwise a new message is sent to the mobile.

TAP is a session based protocol, as opposed to a permanently connected one. Each session comprises a logon, a number of transactions, and a logoff as shown on Figure 15.

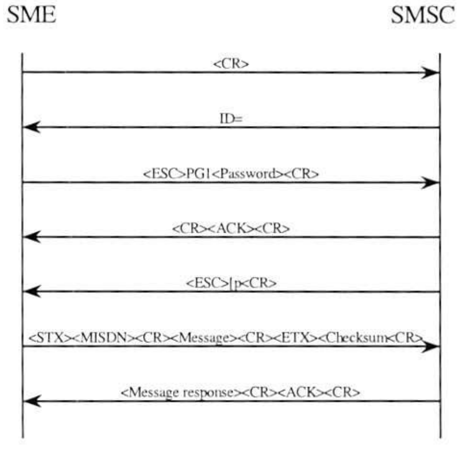

SME
SMSC
<CR>
ID=
<ESC>PG1<Password><CR>
<CR><ACK><CR>
<ESC>[p<CR>
<STX><MISDN><CR><Message><CR><ETX><Checksum<CR>
<Message response><CR><ACK><CR>

Figure 15: Short message submission using TAP

3.4.6. *Other submission protocols*

Many others protocols have been designed and operate over a wide range of hardware interfaces. Most of the protocols can be classified as dumb or smart whether they provide notification of delivery and/or advanced functionality (message deletion, replacement, etc.). To name but a few, SMPP, UCP and CIMD are also popular amongst SMSC manufacturers.

3.5. The SMS evolution

3.5.1. *Compression ETSI 0342*

Breaking the barrier of the 160 maximum characters length of a short message has lead to many research projects in order to develop a compression algorithm. In March 1997, the English company, Vodafone, released details of an ETSI acceptance of their compression algorithm (ETSI 0342) capable of boosting the message length by 50%.

The new algorithm uses a variation of the Huffman coding in which the number of bits allocated to describe a character is inversely proportional to the usage frequency. Therefore a very common letter such as 'e' would be allocated fewer bits than a rather less used letter like 'z'. The encoding and decoding process uses a binary tree as shown on Figure 19:

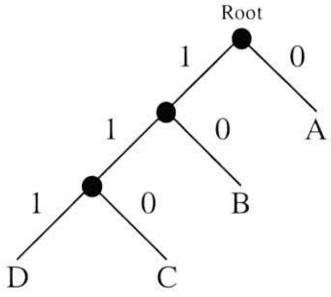


Figure 16: Huffman coding tree

The encoding process of a character starts at the leaf corresponding to the character and goes through the branches collecting the value associated to the branch until the root is reached (e.g. The character 'C' would be encoded as '110'). The decoding process works the tree from the root to the letter. Vodafone's algorithm employs a variation of the Huffman code and a dynamic tree, which is created as the message is being transmitted. The compression process starts with a basic tree to which new characters are added as they appear in the message. To achieve maximum optimisation the starter tree will be based on a language specific sub-set of characters (thus achieving a mere 200 characters per message), as well as a set of dictionaries containing keywords or phrases (such as 'meeting' or 'arrange') to further improve compression. **[Dettmer, 1997].**

3.5.2. *Enhanced Messaging Service*

The Enhanced Messaging Service is the ability to send and receive a combination of formatted text, images, pictures, and melodies to an EMS compliant handset. The EMS message itself can span a number of short messages and uses a combination of headers and binary content to deliver the formatted message. EMS has been built as an enhancement to the existing SMS and does not require any modification to the mobile networks, and still use the signalling channel . Needles to say that an EMS message can only be displayed on an EMS phone. EMS handsets have started to appear in 2001 with Ericsson leading the way. It is still unclear whether EMS will really take off or be superseded fairly quickly by the Multimedia Messaging Service.

3.5.3. Multimedia Messaging Service

This is really the next generation for messaging service. MMS will bring still and moving images with or without sounds as well as rich media text to the end-user. It could be used to send postcards taken using a digital camera connected to a phone using Bluetooth technology or listen to the latest BBC news flash using video streaming. The possibilities are endless. However, the signalling channel can no longer be used to carry the vast amount of information needed and new network components will have to be put into place.

3.6. Conclusion

The Short Message Service is a GSM standard and widely available where GSM coverage is present. It has several unique features which have made it one of the fastest growing messaging services in history:

- A single message can carry up to 160 7-bit characters, enough for two lines of text or 140 bytes of binary data;
- SMS is a store and forward service, it involves the use of an SMSC to act as a storage proxy node between two users;
- Message can be sent and received while voice or data calls are in place as it uses the signalling channel. Even in peak hours when voice calls are hard to place, short messages usually go through;
- SMS is fast, cheap and reliable, messages rarely get lost and are usually delivered within seconds even from one part of the world to another;
- SMS is hugely popular and it unobtrusive nature appreciated by many. It is a fast and convenient way to pass messages.

However, a few disadvantages have in the past limited the acceptance of the Short Message Service. Sending a text message from a phone is frustrating to say the least. It involves pressing a number of keystrokes and prone to typing errors. Until predictive input came along (where a dictionary is used to predict the word a user is typing, thereby simplifying data entry), and e-mail or web gateways were available most people would not even bother using the service. Since then, the growth has been phenomenal and just about every mobile phone user will send at least a couple of messages every month, with some sending hundreds, preferring it to voice call by a long way.
SMS has a few years yet to live and rather than being replaced it will eventually be extended to offer more features and deliver richer content.

4. Code pages and computer systems

4.1. Introduction

A code page is a set of characters specific to a language or hardware platform. It was originally developed by IBM and is now an ISO standard. It consists of a table of characters, numbers, punctuation marks, and other glyphs, and its corresponding numbers in memory that a program uses to display data properly. For a mobile phone, the code pages are installed before the handset is sold. This means that different system will use different extended character sets in the remaining code point in the code page. This can cause a problem in interoperability between mobile phones when sending an SMS message.

There are many different code pages developed by different organisations for different platform and languages to encode the character. Some code pages are 7-bit, others are 8-bit although a 16 bit code set is emerging. Because of the heavy influence of American developers, a lot of software is designed primarily to handle the characters in the Latin-1 alphabet, which covers most of the Western European and American languages. This means that to support other languages, such as the Czech, Slovak, Polish, and Hungarian, it needs to use different code pages. Most of these code pages have the same first 127 characters used in ISO 8859-x standard. Only the remaining characters above 127 vary between code pages. These remaining characters, known as the extended characters set; determine which other languages the code page can support.

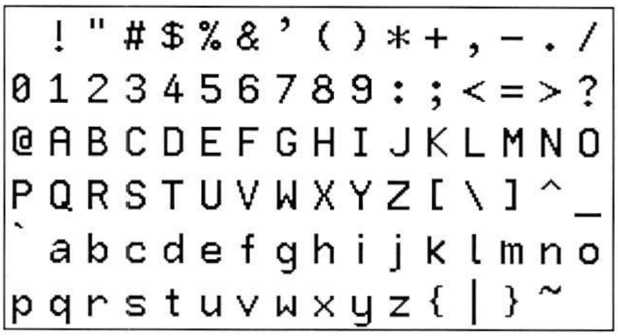
ISO Standard	MS Windows	IBM	Mac	Character set supported
ISO 8859-1	**1252**	**850**	**10000**	**Latin1, Western Europe**
ISO 8859-2	**1250**	**852**	**10029**	**Latin 2, Eastern Europe**
ISO 8859-3	**-**	**853**	**-**	**Latin 3, SE Europe**
ISO 8859-4	**-**	**-**	**-**	**Latin 4**
ISO 8859-5	**1251**	**866**	**10007**	**Latin Cyrillic**
ISO 8859-6	**1256**	**864**	**-**	**Latin Arabic**
ISO 8859-7	**1253**	**869**	**10006**	**Latin Greek**
ISO 8859-8	**1255**	**-**	**-**	**Latin Hebrew**
ISO 8859-9	**1254**	**857**	**-**	**Latin 5 (Turkish)**
ISO 8859-10	**-**	**-**	**-**	**Latin 6**

Figure 17: Examples of different Code pages

Some characters available in one code page of an operating system may not be available in another, and some accented characters may not be represented by the same values across platforms and code pages.

4.1.1. ASCII

American Standard Code for Interchange of Information (ASCII) gives a mapping for 128 characters and has become the standardised ISO 646. These characters consist of upper and lowercase English, American English punctuation, base-10 numbers, and a few controls characters. Although ASCII is very primitive, it is one common denominator contained in all other common character sets and is used in network communication. Therefore, the only means of interchanging data across all major languages without risk of character mapping loss is to use ASCII. But these 128 characters are not enough for most applications and therefore, ASCII has been extended to the 8-bit ISO 8859-x character set, thus allowing an additional 128 characters to be used.

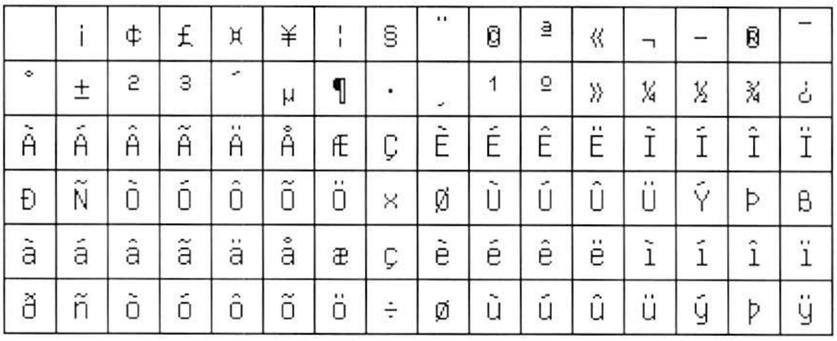


	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

Figure 18: ASCII Character Sets

4.1.2. ISO 8859-x

There are currently 10 different ISO 8859-x standards representing different languages as shown in Figure 17. ISO 8859-1 coded graphic character set is an increasingly popular 8-bit extension of the traditional 7-bit US-ASCII character set. It is already supported by many operating systems including Microsoft Windows. ISO 8859-1 contains graphical characters used in at least 44 countries and can support languages such as Danish, Dutch, English, Finnish, French, German, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish and Swedish. It is already the replacement of the old 7-bit US-ASCII character set and its national ISO 646 variants. In addition, the first 256 characters of the new 16-bit character set ISO 10646/Unicode -that will eventually contain all characters used on this planet- are identical to those used by ISO 8859-1. The extended character set for ISO 8859-1 is shown on Figure 21.

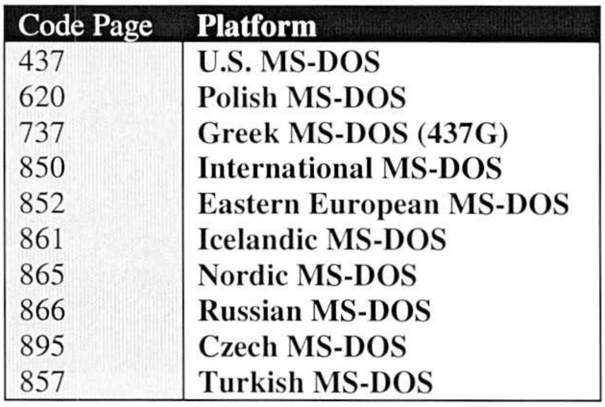


	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Figure 19: The ISO 8859-1 extended Character Set

4.1.3. MS-DOS Code Page

In the MS- DOS operating system, different original equipment manufacturer (OEM) code pages were created so that test-mode computer could display and print line-drawing characters. Figure 20 lists some of the MS-DOS code pages. Some of these OEM code pages are still in used today for direct FAT access and for accessing data files created by MS-DOS based applications. OEM code pages typically have a 3-digit label such as CP 437 for American English. The emphasis with OEM code pages was line-draw characters, which took up a lot of space in the 256-character map leaving very little room for international characters. Different hardware OEMs are allowed to set their own character standards. This can cause character to become scrambled or lost within the same language if two different OEM code pages have different character code points. For example, a few characters were mapped differently between Russian MS-DOS and Russian IBM PC-DOS. So data movement is unreliable and software has to be written to handle each special case.



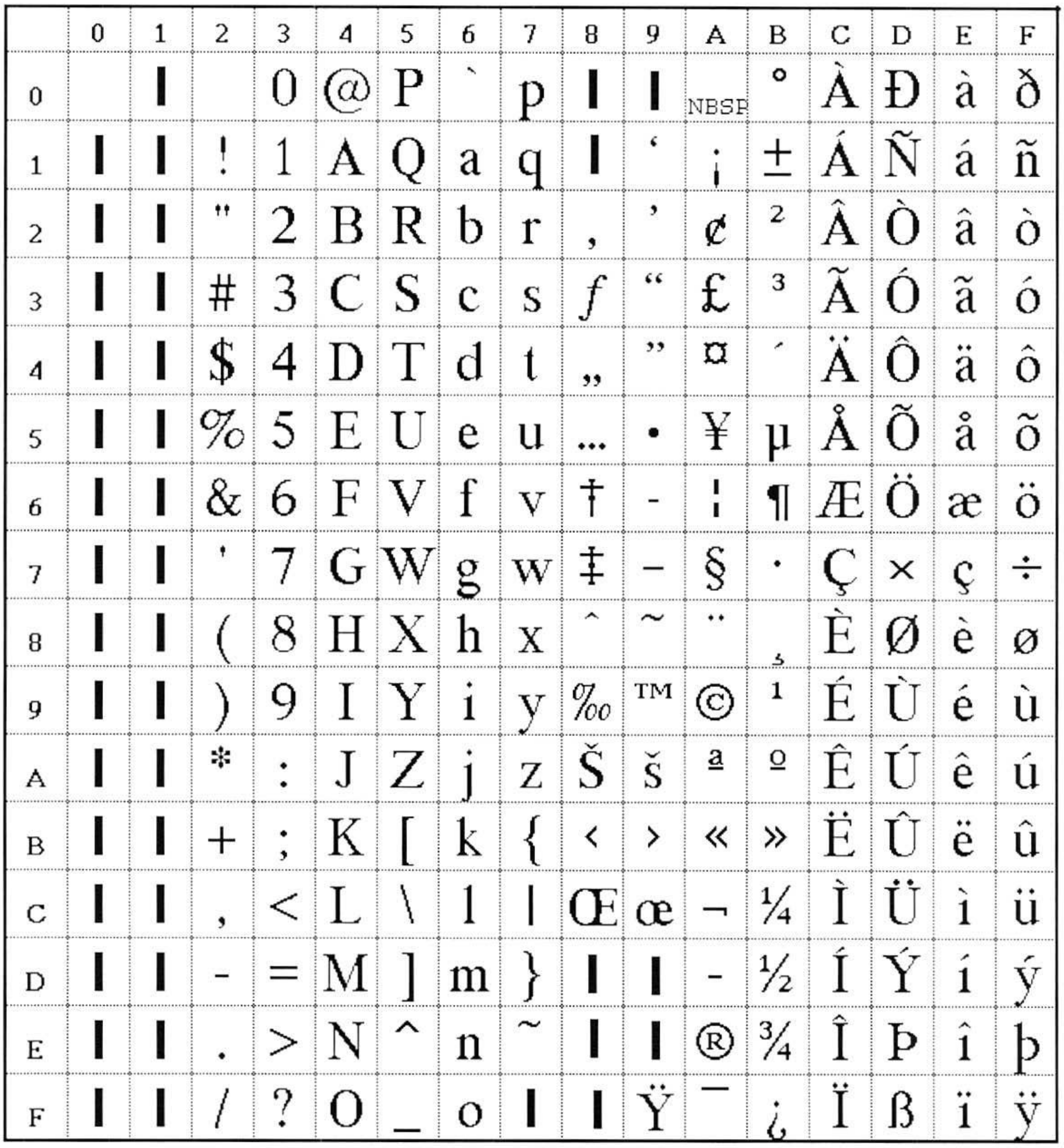
Code Page	Platform
437	**U.S. MS-DOS**
620	**Polish MS-DOS**
737	**Greek MS-DOS (437G)**
850	**International MS-DOS**
852	**Eastern European MS-DOS**
861	**Icelandic MS-DOS**
865	**Nordic MS-DOS**
866	**Russian MS-DOS**
895	**Czech MS-DOS**
857	**Turkish MS-DOS**

Figure 20: Different Code Pages for MS-DOS

4.1.4. Microsoft Windows Code Page

Most developers of international Windows-based programs have been trying to come to grips with character encoding problem. Single-Byte Character Set (SBCS) comes in several flavours, as do the double-byte standards, which are also called multi-byte. Trying to pass data from different character set across networks or between operating systems involves a gauntlet of mappings, and conversions. These code pages continued to pile up with the Russian, Arabic, and Far East versions of Windows. Every time a new language script required character supports, a new code page was created.

Since the Microsoft Windows graphical device interface overrides the need for text-based line drawing characters, the old OEM line drawing character codes could be freed up for international characters and publishing symbols. Windows character sets covers all the 8-bit languages targeted by Windows. It is made up of the first 128 lower ASCII characters code, the 128 upper characters being different for each ANSI character set. These 128 upper characters made up the international characters. One of the Windows code pages is shown on Figure 21.



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0		▮		0	@	P	`	p	▮	▮	NBSP	°	À	Đ	à	ð
1	▮	▮	!	1	A	Q	a	q	▮	‘	¡	±	Á	Ñ	á	ñ
2	▮	▮	"	2	B	R	b	r	‚	’	¢	²	Â	Ò	â	ò
3	▮	▮	#	3	C	S	c	s	ƒ	“	£	³	Ã	Ó	ã	ó
4	▮	▮	$	4	D	T	d	t	„	”	¤	´	Ä	Ô	ä	ô
5	▮	▮	%	5	E	U	e	u	…	•	¥	µ	Å	Õ	å	õ
6	▮	▮	&	6	F	V	f	v	†	-	¦	¶	Æ	Ö	æ	ö
7	▮	▮	'	7	G	W	g	w	‡	–	§	·	Ç	×	ç	÷
8	▮	▮	(	8	H	X	h	x	ˆ	˜	¨	¸	È	Ø	è	ø
9	▮	▮	)	9	I	Y	i	y	‰	™	©	¹	É	Ù	é	ù
A	▮	▮	*	:	J	Z	j	z	Š	š	ª	º	Ê	Ú	ê	ú
B	▮	▮	+	;	K	[	k	{	‹	›	«	»	Ë	Û	ë	û
C	▮	▮	,	<	L	\	l	|	Œ	œ	¬	¼	Ì	Ü	ì	ü
D	▮	▮	-	=	M	]	m	}	▮	▮	-	½	Í	Ý	í	ý
E	▮	▮	.	>	N	^	n	~	▮	▮	®	¾	Î	Þ	î	þ
F	▮	▮	/	?	O	_	o	▮	▮	Ÿ	¯	¿	Ï	ß	ï	ÿ

Figure 21: LATIN 1 ANSI Windows Character Set

4.1.5. Multi-Byte Character Codes

SBCS only provides 256 character codes which is adequate to encode most characters needed for Western Europe but is not enough to represent all characters needed by multi-lingual users in a single font or by the users in the Far East. A Multi-Byte character set (MBCS) is needed to overcome this. MBCS consists of a mixture of Single-Byte and Double-Byte character encoding and provide over 65,000 character codes, more than enough to describe all characters and symbols used in Asian languages.

Language	Character Set	Code Page	Lead Byte	Trail Byte
Chinese (China)	GB 2312-80	CP 936	A1H-A9H B0H-F7H	A1H-FEH
Chinese (Taiwanese)	Big 5	CP 950	A1H-C6H C9H-F9H	40H-FEH
Japanese	Shift JIS	CP 932	81H-9FH EOH-FCH	40H-FCH
Korean	KSC 5601	CP 949	A1H-ACH B0H-C8H	A1H-FEH

Figure 22: Windows Code Pages for Far East Version

In Double-Byte Character Set (DBCS), each character consists of a lead-byte plus a trail-byte. These two bytes must be treated as a unit when it is read. Randomly accessing strings, scanning for special delimiters or performing case conversions can produce unexpected results if a lead-byte or trail-byte is treated as a single byte character. Special care has to be taken when dealing with MBCS. Strings must be scanned linearly from the beginning to the end. A double byte character should be noted as they are found. Insertion, deletion, truncation, and cursor movement must never separate the lead-byte with its trail-byte. For this reason, string pointer should not be incremented or decrement by -- or ++ as in C. Instead, a macro or function calls, which deals with DBCS should be used.

4.2. SMS Code pages

The Short Message Service (SMS) used in mobile phones also has its own code page. This code page can either be in 7-bit, 8-bit or 16-bit . All mobile station should have the ability to store a short message coded in any alphabet. It is usually installed when the mobile phone was manufactured. Some mobile phones can support up to 7 different European Languages while mobile phone in the Far East can support Chinese and Thai. Nokia has launched the first handset capable of displaying Far East characters. The handset uses a 16-bit character set and has a bigger and wider display.

The default character set for SMS is shown in Figure 23 **[ETSI 3.38 1996]**. This default alphabet is mandatory and should be supported by all MSC and SMSC. The character is 7 bit long and contains all the English Alphabet used today and some other languages like Greek and French. It also contains three control characters, namely: Space, Carriage Return and Line Feed, numbers and a few symbols. The code point marked by 1) indicates that it is not in used and will be displayed a space.

				7	0	0	0	0	1	1	1	1
				6	0	0	1	1	0	0	1	1
				5	0	1	0	1	0	1	0	1
4	3	2	1		0	1	2	3	4	5	6	7
0	0	0	0	0	@	Δ	SP	0	¡	P	¿	p
0	0	0	1	1	£	_	!	1	A	Q	a	q
0	0	1	0	2	$	Φ	"	2	B	R	b	r
0	0	1	1	3	¥	Γ	#	3	C	S	c	s
0	1	0	0	4	è	Λ	¤	4	D	T	d	t
0	1	0	1	5	é	Ω	%	5	E	U	e	u
0	1	1	0	6	ù	Π	&	6	F	V	f	v
0	1	1	1	7	ì	Ψ	'	7	G	W	g	w
1	0	0	0	8	ò	Σ	(	8	H	X	h	x
1	0	0	1	9	Ç	Θ	)	9	I	Y	i	y
1	0	1	0	A	LF	Ξ	*	:	J	Z	j	z
1	0	1	1	B	Ø	1)	+	;	K	Ä	k	ä
1	1	0	0	C	ø	Æ	,	<	L	Ö	l	ö
1	1	0	1	D	CR	æ	-	=	M	Ñ	m	ñ
1	1	1	0	E	Å	ß	.	>	N	Ü	n	ü
1	1	1	1	F	å	É	/	?	O	§	o	à

Figure 23: Character Table for GSM Default 7 bit Alphabet

4.2.1. SMS Character Coding Scheme

When a SMS is sent from an MS to another MS, the character coding scheme used by the message is indicated in the TP-Data-Coding-Scheme field (TP-DCS) of the TPDU **[ETSI 3.41 1996]**. As explained in a previous chapter, the TP-DCS field is 1 octet long. The octet is used according to a coding group, which is indicated in bits 7 to 4, as described in Figure 24. When the TP-DCS bits 7 and 6 are set to 00, it indicates a general coding scheme. The bit 5 will indicate if the text is compressed or not using the standard GSM compression algorithm. The algorithm has recently been approved by ETSI and is described in section 3.5.1. Bit 4 is used to indicate if bits 1 and 0 are reserved, or have a message class meaning. There are 4 classes used to indicate how a message should be treated. Class 0 message, marked by 00 in bits 1 and 2, will have the message displayed immediately and send the acknowledgement to the SMSC. Class 1, marked by 01, will have the message stored in the MS and Class 2, marked by 10, will have the message stored in the SIM. Class 3, marked by 11, will send the acknowledgement to the SMSC when the message has been received by the MS.

Bits 7 to 4	Description
00xx	**General Coding**
0100-1011	**Reserved Coding**
1100,1101,1110	**Message Waiting**
1111	**Data coding/message class**

Figure 24: Character Coding Scheme for bit 7 to 4 of the TP-DCS

The type of character set used in the TP-UD is indicated in bit 3 and 2 of the TP-DCS. When a default character set is used, it is set to 00. An 8 bit user defined character is indicated by 01, 16 bits UCS character is indicated by 10 and 11 is reserved for future use.

As for the remaining coding scheme, 1111 of bit 7 to 4 represent a data coding or message class. Bit 2 of that coding scheme will be used to indicate if the message is using a default 8-bit character. The remaining bits remain unchanged. As for the message waiting encoding group, 1100 allows the mobile phone to discard a message. 1101 and 1110 allows an indicator to be displayed as an icon about the type of message waiting on the systems. The text included in the user data is coded using the default character and the originator address may be obtained. 1110 will have the text coded in the uncompressed UCS-2 alphabet. Bit 3 is used to set the indication active or inactive and bit 2 is reserved. The actual message type will be encoded in bit 1 and 2. 00 indicates that a voice mail is waiting, 01 indicate a fax, 10 indicate an Email, and 11 is reserved for future use.

4.3. Unicode Standard

Unicode is the world-wide character-encoding standard destined to replace ASCII and the multitude of other single and multi-byte character sets currently in existence. It started out as an informal collaboration between engineering teams at Apple and Xerox to overcome the limitations imposed by incompatible computer standard. This collaboration was later joined by other major computer companies eventually becomes the Unicode Technical Committee from the Unicode Consortium which is a non-profit organisation founded in January 1991. It is also a subset of the ISO 10646 standard. The 64Kbytes encoding layout is shown on Figure 25. This expansion provides codes for more than 65,000 characters, of which 29,000 unused code are reserved for future use.

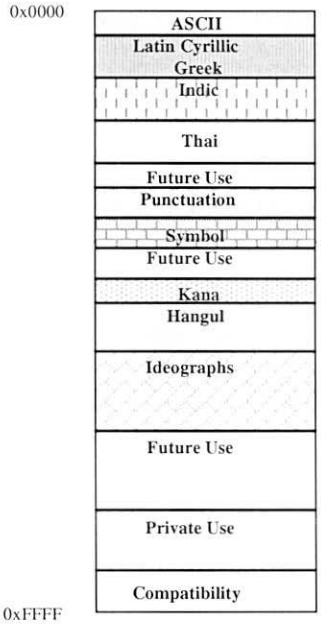


Figure 25: The Unicode Encoding Layout

The Unicode encoding provides the capability to encode all the characters used by written languages throughout the world. This includes the Latin alphabet used by English, the Cyrillic alphabet used for Russian, the Greek, Hebrew, Arabic, and other alphabets used in countries across all continents. The largest part of the Unicode standard is devoted to thousands of unified character codes for Chinese, Japanese, and Korean ideographs which have used the Han Unification. It also supports symbols, with codes defined for punctuation marks, mathematical symbols, technical symbols, arrows, and others. It encodes each character and vowel marks separately, and allows characters to be combined to create a marked character. In addition, the Unicode standard reserves over 6,000 code numbers for private use, which software and hardware developers can assign internally for their own characters and symbols. To keep character coding simple and efficient, the Unicode standard assigns each character a unique 16-bit value, and does not use complex modes or escape codes to specify modified characters or special cases. This simplicity and efficiency makes it easy for computers and software to handle Unicode-encoding text file **[KANO(i) 1995]**.

Although there is a minor penalty for storing uniform 2-byte values for each character, Unicode has significant advantages. It eliminates the confusion of overlapping, single-byte (8-bit) code pages in which a character's identity is dependent on the active code page. And also, the uniform 16-bit width of each character makes it easy to determine character boundaries in contrast to multi-byte character

4.4. Problems with code pages

In the SMS code page, the default alphabet set only supports the ASCII character with some French and .Greek characters. Support for other character will require the mobile company to define the 8-bit character set. Every mobile phone from one company will have its own extended character set defined in its code page. Since the system will be using its own extended character set and data is send using the code point, it might not be able to exchange messages between different service providers without corrupting the information. The system will receive the character value in the message and display the corresponding character in its own character set For example, the character ~, which is not available in the default code page, will be assigned in the extended character set. One service provider might assign this character at a different code point from another service provider. Therefore, when a person sends a message to that mobile, the ~ character might be replace by another character based on the code point numerical value. This will corrupt the message. Sending SMS to a computer or vice versa can also be very difficult because the code page used in computer is different from the SMS code page.

4.4.1. Number of characters available

The number of characters available depends on the number of bits used by each character in the code page. In the default 7 bit SMS code page, only 128 characters are available to represent the English, French and a few Greek characters. If a service provider uses the extended characters, the maximum number of characters available will increase to 256. The service provider will then have to decide which characters to include. This will vary between service providers. The 256 characters may be enough to assign most of the characters used by three typical countries in Europe. But to choose which languages to be assigned is not easy. This can be a limitation for a user sending an international SMS message. For example, a service provider might have assigned all the characters based on the ISO 8859-1, and another service provider might have assigned a different character set based on the ISO 8859-9. When the message is exchanged, some of the characters may be missing from the receiver code page since ISO 8859-9 supports Turkish while ISO 8859-1 supports Icelandic. This can cause the message to be displayed incorrectly due to the unavailable characters. Unless a 16 bit character set is used, the number of character available will be very limited. This 16-bit character set will contain a 256 by 256 matrix of characters. This is more than enough to encode all European languages but will be usually defined by the service provider making inter-operability even more difficult.

4.4.2. Message Size of 140 octets

SMS only allows 140 octets of data to be transmitted in its TPDU. This imposed a constraint in the number of characters that can be sent at one time. The number of characters depends on the number of bit of the code page used. Code page can be designed using a 7-bit character set, 8-bit character set, and 16-bit character set. If a message is using an 8-bit character set, then the number of characters allowed to be sent is 140 characters.

This is calculated using the formula:

$$NumberofCharacters = 8 \times \frac{MessageSize}{NumberofBits}$$

This means that if a 16 bit character set is used, which can accommodate a double byte character set, the number of characters that can be sent decreases to 70 characters. The service provider will have to make a compromise between the number of characters available and the number of characters allowed to be sent, unless an alternative solution such as a compression algorithm is used.

4.4.3. Network Incompatibility

When a mobile sends a message to another mobile, the GSM network can transmit the message without corrupting the message. But, if the message has to be sent to a computer, the message can be corrupted. This is because the message will be sent to a Local Area Network (LAN) which imposes a 7-bit rule. The LAN has a formal rule in which certain types of messages, in particular email, can only use a limited number of 90 basic characters. Any documents that contain octet values outside the 7-bit range need a content transfer encoding applied before transmission over certain transport protocols **[Freed et. Al 1996]**. There are several solutions to this. The first is to accept it and try to squeeze the required characters into the restricted set by replacing the English letters or standard symbols. The second is to ignore it and just send the message and hope it will arrive uncorrupted. The third is to change the network by adopting an extended network system that properly negotiates a full 8 bit. This solution is not common, as it is not done at user level but by a network administrator. The final solution is to encode the characters so that a full range of extended characters can be transmitted over the 7-bit communication line. This is the most suitable and useful solution but the message has to be decoded by email software. This kind of encoding is part of the larger specification for email software called Multi-purpose Internet Mail Extension (MIME) **[RFC 1521]**.

MIME is a specification that any email program should adhere to. It informs the receiver of the content of the email. The external SMSC which is connected to a LAN, has to implement the MIME encoding at the application layer to prevent characters, such as "{ }[]\|~_", to become corrupted. The message received from the GSM network has to be encoded before it is sent to the LAN. MIME supports all the ISO 8859-x character set standards. It is also supported by almost every transport protocol used today. Using the MIME system, a standardised line can be added to the email's header saying which character set it is using **[RFC 2110]**. This will allow the end system application to use the code page to decode the message.

4.4.4. Keyboard to Code Page Mapping

This keyboard to code page mapping is more of an issue that needs to be looked at rather than a problem to be solved. Users should be able to switch between all available languages and corresponding keyboard layouts configured on the system. When a key is pressed on the keypad or keyboard, a signal is send to the processing unit of the system. The processing unit then gets the code point for that particular character based on the signal sent. There will usually be software to map the signal from the keyboard to the code point on the code page. For example, in Windows 95, when the key 'A' is pressed, the keyboard sends a location signal 'xxyy'. This 'xxyy' will be checked against the keymap and the code point value '#1' is sent to the screen and network. This means that if a system has to support international code pages, the keymap software should be able to adapt itself to the new code page. The same requirement is also needed in the mobile phone to map the keypad on the mobile phone to the code page.

To enter a character that is not represented on the keypad, it will require a user to use more than one key. When this occurs, the system should be able to detect the last keypad that will represent the character based on the code page used. For example, when a user enters "control 1", it will be mapped to a certain character. But when it is followed by another key, then the system must be able to know if that key will represent a new character or has to be encoded with the previous two characters. There is still also no efficient way to input the Chinese language in the mobile phone. Users are required to phone their own paging service and ask to be paged with the message before it can be send to another phone. In Far East editions of Microsoft Windows, is included a built-in input method editor (IME), which will automatically return the correct characters to any DBCS-enabled or Unicode-enabled application. Windows also provides a full bi-directional text layout support.

4.4.5. User Interface

User interface is another issue that needs to be looked at when dealing with internationalisation. When a message is received on the MS, the MS should be able to display the message. The display panel should display all the characters used in the message whether it is a DBCS or SBCS based on the code point of the code page. Nokia's latest offer to the crowded cellular phone market includes a Chinese user interface with a screen that can hold two rows of six simplified or traditional characters at once, or four rows of 13 letters. It features a large dot matrix screen, which is ideal for SMS messages. The messages can run as long as 70 Chinese characters or 160 letters. A display driver will be required to map the character to the code point and display it on the screen. Displaying the character in the correct font is also very important, especially in Unicode.

4.5. Possible Solutions

4.5.1. Character Sets Mapping

This is the easiest to implement to deal with the code page compatibility problem. It will be used as the solution for this project. A system is usually allowed to use only one code page at a time. When a system is set to use a particular code page, it will not be possible to change one code page to another without reconfiguring and restarting the system. Therefore, an alternative is to change the code point used by the character. Character set mapping is procedure which will convert a code point assigned to a character to another code point based on the code pages used by the systems **[ETSI 7.05.1996].**

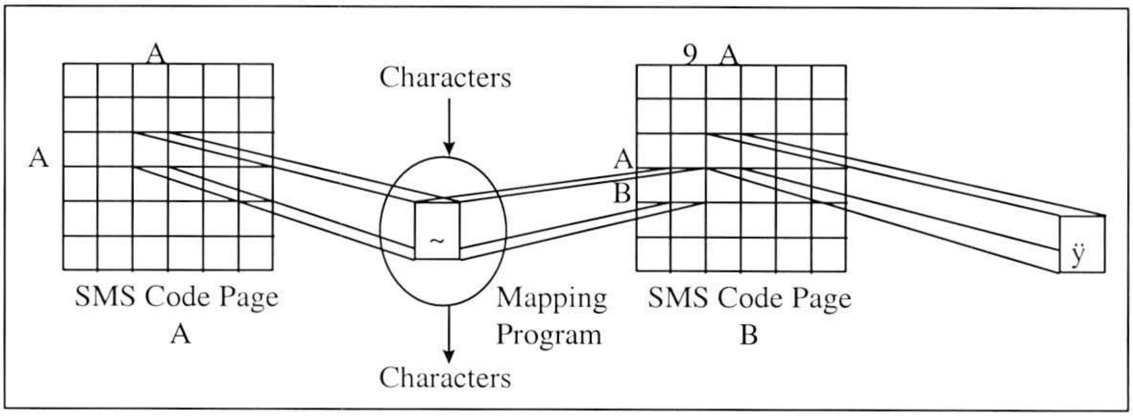


Figure 26: Two codepages representation of the character ~.

To illustrate the procedure, assume that user A is using code page A and user B is using code page B. When user A sends a message to user B, characters such as ~ may be corrupted due to different code point representation. As shown in Figure 26, the character ~ may be represented by AA in A and 9B in B. If mobile phone B has displayed the character using its code page, the character ÿ at code point AA will be displayed. A program will be required to map the character set A to character set B. Before that program can perform the mapping, the program will need to know the code pages used by systems A and B. They have to be made available so that a mapping relationship between code page A and code page B can be obtained. From there, the program will use that information to convert the code point used by a character to the local code point. This solution is usually implemented in the end system application layer or at an SMSC. This is because both code pages can be made available easily. If a solution is implemented in the end system, the user will not be restricted to the code page set by the SMSC. But, this will require some sort of database to store the mapping relationship, which can take up some space in a mobile phone.

4.5.2. Reference By Name

This solution is used to overcome the limited number of character available in the code pages. It uses an entity to represent a character that is not available on the standard ASCII character set. One character, such as #, will be used as a start point marker to the entity reference. Another character, such as *, will be used as an end point marker to the entity reference. For example, to represent the character Æ, it is denoted as #AElig*, short for A-E Ligature. With this solution, the dependency to other code pages is removed because the entity name allows the SMS to represent the character that is not in the default code page. Hence, every system will be able to use the same representation. With the 128 characters available, the range of entity names is unlimited.

The problem with this solution is that the entity has be given a name which is meaningful and can be understood by anyone. For example, #AElig*, people who understand English will know that it represents A-E ligature. But people who cannot understand English will have no clue what it means or represents.

4.5.3. Standardisation

This is the mother of solution to all code page problems. If every system had used only one standardised code page, then the incompatibility problem between systems would not exist at all. The missing character problem and limited characters are also eradicated as the user will be using the same code page. If more characters are required, the standard can be updated by including that character. One example of such standardises character set which represents all the characters used in the world, is Unicode as described earlier.

However, to agree on one standardised SMS code page is an expensive and difficult task. Even if a standardised code page had been created today, mobile company would be reluctant to change their current code pages. It will be very expensive for them to change their existing code pages. And also, if a standardised code page is to be used, the old system that uses the old code pages will not be able to communicate with the new system. The company could have used Unicode for their SMS messaging from the beginning. But the number of characters, that are allowed to be sent, has been the obstruction for the company to use it. Currently, there are no standardised compression algorithms defined by ETSI to compress Unicode.

5. Standard for electronic mail delivery

5.1. Introduction

Since its appearance on the early 1960, electronic mail message transfer has become the most widely used application service of computer networks. The specification quickly evolved and in 1982, the ARPANET proposed the Simple Mail Transfer Protocol (SMTP) [Postel J.B, 1982] to allow delivery of messages across computer networks. The specification quickly evolved to include non textual data in the form of attachments and the MIME extension was proposed [Freed et. al. 1996].

The wide availability of electronic mail client software and its increasing popularity made it a strong candidate as a starting point for an access method to the GSM Short Message Service.

This chapter introduces the SMTP specification and highlights the MIME extensions.

5.2. Electronic mail formats

The SMTP is an ASCII based mail delivery protocol. RFC822 [Crocker, D. 1982] specifies the format of the email messages. The message consists of a basic envelope followed by a blank line then the message body. Table 5 lists the principal fields of an electronic mail header:

Table 5: RFC822 e-mail header fields

Field	Description
From:	**Identifies the sender**
To:	**Identifies the primary recipient(s)**
Cc:	**Identifies the secondary recipient(s)**
Bcc:	**Identifies the Blind carbon copies**
Subject:	**Summary of the email content**
Message-ID	**Unique identifier assigned by the originating SMTP server**
Received	**Routing information added by every hop, e.g. SMTP server receiving and forwarding emails.**
X-	**Custom information**

The routing information contained in the email header is very valuable to identify the number of hops (or SMTP server), latency and ID.

The local delivery agent may add other headers to provide information about the time and date of receipt, and a local unique identifier for the message. The body of the message follows. Initially electronic messages were mainly written in plain English using the ASCII character set. This was a major problem when it came to sending language specific accentuated characters such as German or French, or supporting non-Latin alphabets such as Russian. Also, RFC822 did not provide any mechanism to send messages not containing any human readable data (e.g. binary files). As a result the Multipurpose Internet Mail Extensions (MIME) [Freed et. al. 1996] was proposed in RFC 1341 and updated in RFC 1521 and is now widely used. MIME added a set of rules defining how to send non-ASCII messages. Five new headers were created and are listed on Table 6.

Table 6: Additional headers added by MIME

Header	Description
MIME-Version:	**Identify the MIME version**
Content-Description:	**Human readable description**
Content-ID:	**Unique identifier**
Content-Transfer-Encoding:	**Information on the body encoding**
Content-Type:	**Information on the body structure**
boundary	**Separator for each individual attachments**

The *MIME-Version* header is the most important as the lack of it would be interpreted as the message being in plain text English and treated as such. The *Content-Transfer-Encoding* gives information on how the body of the message is encoded and has to be processed. Five encoding schemes are provided with ASCII and Quoted Printable being the most relevant to this paper, the other encoding types apply to binary data and are not used to encode text. The ASCII scheme is just plain 7-bit ASCII and comprises 128 characters. The quoted printable encoding scheme provides a way to describe 8-bit characters by replacing each character with a value above 127 with an equal sign followed by the characters value as two hexadecimal digits. For example a character with a value of 255 would be represented as '*=FF*'.

Table 7: RFC1521 MIME types and subtypes

Type	Subtype	Description
Text	**Plain**	**Unformatted ASCII text**
	Richtext	**Text including simple formatting commands**
	Html	**Text described using HTML tags**
Message	**Rfc822**	**An RFC822 message**
Multipart	**Alternative**	**Same message expressed in different format**
	Mixed	**Independent parts**

The *Content-Type* header defines the structure of the body of the message as a type and a subtype separated by a slash. The relevant types to this paper are depicted on For example a character with a value of 255 would be represented as '*=FF*'.
Table 7 (e.g. thus describing text-encoded messages). Many other types and subtypes exist and new ones are regularly added.

The *Text* type splits into three subtypes describing plain text encoding or text encoded using a simple set of rules. The *Richtext* subtype uses tags to provide machine independent encoding of basic formatting such as bold or Italics. The *Html* subtype extends the formatting possibilities using the Hyper Text Mark-up Language (HTML). When text is encoded as *Richtext* or *Html* the character set used by the sender's email client is specified as a *charset=* additional field. MIME currently supports all ISO8859 character sets.

The *Multipart* type allows the body of the message to contain more than one part, each separated by clearly defined boundaries. The *mixed* subtype allows each part of the body to be different, with no additional structure imposed. The *alternative* subtype, however, allows the same message to be expressed in different format, with each part ordered from the least complex (e.g. plain text then rich text then HTML). Each part of the body will then be described using the same type and subtypes.

The task of extracting the relevant information (i.e. textual only, the human readable one) and sending it to a mobile user as a short message is then rather complex considering all the different formatting types available. More importantly, the lack of restrictions in the standard means that software vendors have taken liberties in implementing their client software and that the extraction of the data is far more difficult than it should be.

5.3. Conclusion

The SMTP is widely accepted as the de facto standard for electronic messaging in the Internet world. From a simple text based specification in the early 60s, the standard has evolved to include additional character sets and to add the ability to transport non textual information in the form of attachments.

In a similar manner to the Short Message Service, SMTP uses gateways as store and forward nodes to carry the electronic mail from originator to recipient. The evolution of the Short Message Service is also very similar and one could compare EMS with the addition of MIME to the SMTP specification.

With the generalisation of cheap and easy Internet access email is also now not only reserved to the office worker but available to a wide range of users of all ages. It only seemed natural to want to interface it with the Short Message Service in the most seamless possible manner.

6. Gateway design and implementation

6.1. Introduction

This section describes a gateway service using SMTP as an access point for the short message service. The choice for SMTP came from a simple conclusion that a versatile and robust protocol would lead to less development in order to develop a working version of the gateway. What's more, using SMTP meant that client application was widely available, as any common e-mail package would be suitable to submit a short message. Other access points (SOAP, XML, HTTP based) have since then been implemented to cope with larger amounts of messages or offer a wider range of submission methods. They are however not covered in this document.

6.2. Protocol architecture

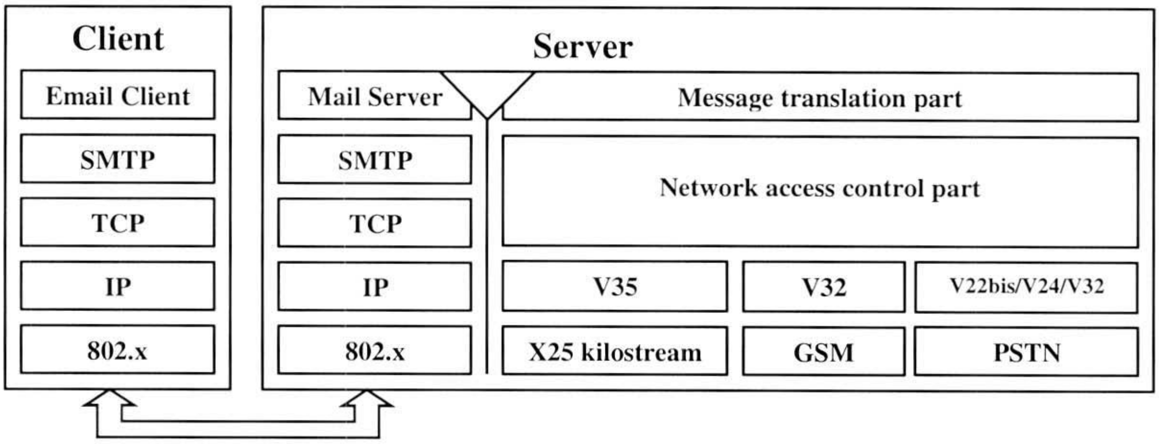


Figure 27: gateway protocol architecture

Figure 27 describes the protocol architecture of the gateway, all communications devices are connected to the gateway using a standardised serial interface set at different speeds (e.g. V22bis/V24/V32/V35). The gateway operates at the application layer; the mail server is configured to deliver specific email addresses to the message translation part as opposed to a local mailbox. We use the alias features provided in Sendmail to map an email address to the input of a process. Consequently every time a new e-mail is received by the gateway, the mail server starts a new instance of the message translation process, thus allowing parallel processing of incoming e-mails. Communication between the Message Translation Part (MTP) and the Network Access Control Part (NACP) is achieved using UNIX inter-process communication (i.e. FIFO or signals) allowing interrupt driven operation and minimising processor usage.

6.2.1. The SMTP interface

One of the rules of software development is "reusability". Indeed, it seems pointless to implement, even partly, a SMTP stack if one is already widely available. Development time is shortened and you then benefit from a product that is both extensively tested and supported. A quick survey positioned Sendmail as the most likely candidate: it is used by very large corporations to handle enormous amounts of e-mail and has a reputation for robustness and scalability. The configuration is slightly more complex than for "normal users", as incoming e-mails are not delivered to a mailbox (a flat text file in essence) but to the MTP of the gateway.

As in every concurrent server, Sendmail consists of a master daemon listening for incoming email and creating child to serve the incoming request. The child then processes the incoming email, mainly adding additional fields to the header, and then proceeds to final delivery. An instance of the MTP is created and the content of the email is passed from the child to the MTP through a standard UNIX pipe. Finally the child waits for the return status of the MTP to log the transaction and exit. Unfortunately, this approach has a major drawback as a new instance of the MTP has to be initialised and loaded into memory for every incoming email.

6.3. The Message Translation Part

The MTP Part is divided into six sub-layers, each taking care of one step in the overall short message creation process. The six sub-layers are depicted in **Figure 28**.

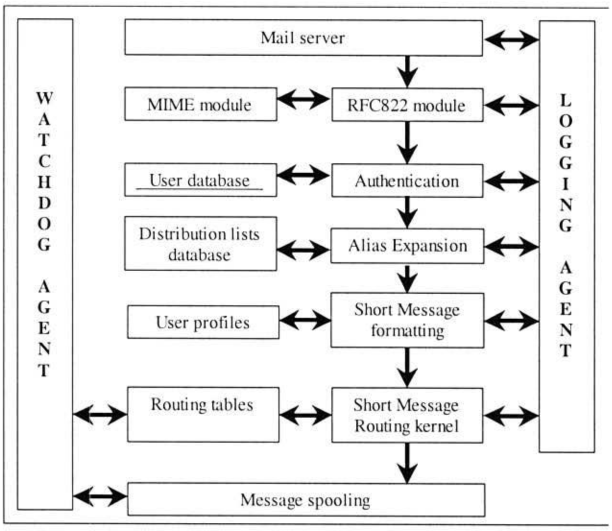


Figure 28 : Message translation part sub-layers for Short Message creation

6.3.1. The RFC822 and MIME modules

The ***RFC822 module*** processes the incoming email and extracts the relevant information from the header. If necessary, the ***MIME module*** is loaded to decode the body of the message, discarding attachments, and removing non-textual information and formatting tags when applicable. The module looks for recipients in the *From:, received, To:,* and *CC:* headers fields; it also extracts the *Subject* header and passes the information to the next sub-layer. Optional parameters can also be passed on in the email header, subject or body (see API in appendix for more details).

6.3.2. The authentication module

The message ***Authentication*** part provides security to the system by restricting access to registered users. The authentication procedure is fairly basic and will check for the existence of an account and user and then proceeds with checking a password if one has been defined in the user profile.

Registration details are kept in a database and used for billing, authenticating and message formatting. The structure is similar to a standard UNIX password file, in which each field separated by ':' An example is shown on Figure 29.

```
Dialogue:Guillaume Peersman:3::pageguiom:447976918568:%R%Sp(%S20)%Sp%B%Ne:905255158:::FOC
```

Figure 29: User profile example

The user profile is a record of the following information:

Field	Description
Account name	**The Account**
User name	**The user within the account**
User type	**For billing purposes**
Password	**Stored in encrypted format**
User	**The alias for SMTP access**
User's number	**The user's MISDN number**
Message format	**The default message format**
Registration date	**in epoch format**
Administrator's details	**For billing purposes**
Group name	**For billing purposes**
Product code	**For billing purposes**

While the in-depth description of each of the fields is not relevant to this document, the user type is probably the most significant. It controls what the user can do with his account.

Three main types of users are defined:

1. receiver user
2. sender user
3. distribution list (a sort of closed user group version of a sender user)

The receiver user is a one to one relationship between an email address (the alias@domain) and an MISDN number (the user's number). In this case the user (or anyone else) is only able to send messages to his own phone.

The sender user is a one to many relationship between the email address used and a list of MISDN numbers passed on with the request. In this case the body of the email is used to store a list of recipient as well as the message itself.

The distribution list defines another one to many relationship but with limitations. The list of recipient is static and stored on the server.

The authentication feature can be disabled at will on a per user basis. When enabled, the authentication sub-layer encrypts the *Subject* of the e-mail using the standard UNIX *crypt* function and compares the resulting encrypted password with the one from the user profile.

6.3.3. The alias expansion module

This module simply expands receiver users and distribution lists in a list of MISDN numbers. Note that a distribution list can contain other receiver users and that a sender user can include distribution lists and/or receiver users in his recipient list.

6.3.4. Message formatting

At this stage, the relevant information has been extracted from the email, the recipient(s) identified and their corresponding MISDN routed. The user profile is now used to format the short messages prior to submission. The message format is defined by a string of keywords destined to be substituted by their values. Some of the keywords have no corresponding value but merely set options which have implications on the formatting or message submission properties.
Every user account on the server is assigned a default message format. The format defines which fields in the message should be included in the short message(s) and a number of options. The Message format consists of a succession of keywords, replaced by their corresponding values when creating one or more short messages as depicted in Table 8 and .Table 9.

Table 8: Keywords for message format control

Keywords	Corresponding fields
%Sp	**A space character (ASCII 0x20)**
%CR	**A carriage return (ASCII 0x0D)**
%LF	**A line feed (ASCII 0x0A)**
%Sn	**The first n characters of the subject field.**
%B	**The email body**
%F	**The originator name if available**
%R	**The originator email address**
%D	**The current date**
%T	**The current time**

Note that if *n* is omitted then the limit will only be the maximum payload of a short message (e.g. 160 7-bits characters).

Table 9: Keywords for message options control

Keywords	Corresponding options
%Smtp_Auth	**Authenticate against the originating SMTP server**
%Ne	**Notify Originator or Administrator of errors**
%Ns	**Notify Originator or Administraton of submission to the GSM network**
%Nd	**Notify Originator or Administrator of delivery to the mobile phone**
%Scsm	**The mobile terminal supports short message concatenation as defined in [GSM 03.40].**
%Ucn	**Submit up to n (possibly concatenated if supported) short messages**
%Pn	**Set message priority to 'n' (0 to 9). Can not be overridden**
%OA	**Allow substitution of the originating address of the outbound message.**

As an example, consider the following message format:

%R%Sp(%S20)%Sp%B%Ne%Ns%Nd

Composed of 5 keywords and 3 options.

After stripping out the options, only the keywords are left:

%R%Sp(%S20)%Sp%B

After replacing the keywords the short message is formatted:

user@domain (subject) body

The position of the keywords is free and additional characters (such as the round brackets above) can be included. For example the same email will generate a completely different short message using the same options and keywords and the following format:

(%S20)%Sp%B%Spfrom%Sp%R%Ne%Ns%Nd

will generate the following short message:

(subject) body from user@domain

6.3.5. Message routing

This section describes how messages are routed to their home GSM network. The routing is very similar to the one used for local IP routing and this will be used as an analogy to illustrate the algorithm. Let us consider a small private local area network (LAN) using a class C address of 192.168.2.x. The LAN is divided into 6 sub-networks, each containing

30 hosts by using 255.255.255.224 as the netmask (see Table 10). The host number 17 in the sub-network 5 has an address of 192.168.2.177.
Every time a packet is transmitted from that host on the network, the destination address is used to route the packet. The destination address is bitwise ANDed with the netmask and the result compared with the sender's sub-network; if they match then the computer should be directly reachable, otherwise the packet is forwarded to the default gateway. The default gateway then attempts to route the packet to one of the networks it is directly connected to or forward it to another gateway. The process is repeated until the packet finally reaches its destination or expires on the network. Table 10 describes how the local sub-networks and hosts are extracted from the IP address. In this example any packet leaving host 17 (IP address of 192.168.2.177) with a destination address in the range 192.168.2.1 to 192.168.2.254 is directly reachable, even if some hosts will be on a different sub-network.

Table 10: dotted decimal and binary representation of IP address and submask

	Dotted decimal	Binary representation
IP address	192.168.2.177	1100 0000 . 1010 1000 . 0000 0010 . 1011 0001
Netmask	255.255.255.224	1111 1111 . 1111 1111 . 1111 1111 . 1110 0000
Bitwise AND	192.168.2.160	1100 0000 . 1010 1000 . 0000 0010 . 1010 0000
Subnet number	5	0000 0000 . 0000 0000 . 0000 0000 . 101
Host number	17	0000 0000 . 0000 0000 . 0000 0000 . 0001 0001

The same concept of local network and default gateway is used to route short message on the gateway described in this section. The local network is defined as the country the gateway is in (i.e. the UK) and the default gateway is one of the GSM or Paging networks SMSC (This assumes connection to all UK based GSM and Paging networks). Every time a message is submitted from the gateway, the destination address (i.e. a mobile/pager number) is used to route the message. If the number matches the local country code (i.e. starts by 44 for the UK, 33 for France, etc.) then the corresponding GSM network's SMSC should be directly reachable; otherwise the message should be submitted to the default SMSC (one of the UK based SMSC). The default SMSC is then given the responsibility to route the message to the recipient (mobile station or pager). The ability of the default SMSC to forward the message to its final destination mainly depends on roaming agreements and configuration of the many SCCP switches along the way.

Table 11: IP numbering and MISDN numbering analogy

	IP	MISDN
Full address	192.168.2.177	44.7976.123456
Network part	192.168.2.160	44.
Netmask	255.255.255.224	44.sub-network list.any
Sub-network part	5	7976
Hosts number	17	123456

To further the analogy, we can represent an example mobile number of 447976123456 as 44.7976.123456 and split it into the country code of 44 (e.g. UK), the network number of 7976 (e.g. Orange) and the host number of 123456. Splitting an MISDN number into network, sub-network and host part is a bit more complex than for an IP address. The notion of netmask is slightly different as all the part are of variable length: the country codes range from 1 to 3 digits, while some sub-network addresses can vary from 3 to 5 digits. While IP routing uses a bit-wise masking operation, regular expression matching is used instead. The network part (country code) is already known and trivial to extract. The sub-network part however relies on the relevant authorities (e.g. OFTEL in the UK, ART in France) to disclose the complete list of allocated addresses and corresponding sub-networks.

Using the list, routing is achieved by attempting to match an element of the list with the left hand side (LHS) of the mobile number: 7976 matches 7976123456 using the following regular expression:

7976123456 =~ /^7976/

This process has to be repeated for every known sub-network address until a match is found. Once the sub-network is known, messages are submitted using the current settings (i.e. interface) for that sub-network. In practice the prefix list is stored as a binary tree and searches are therefore extremely fast. The next sections describe a definition language designed to easily configure and maintain the routing tables of the gateway and address issues such as scalability, least cost routing and resilience.

6.3.5.1. The network definition language

The United Kingdom alone counts 4 GSM networks (6 with the Isle of Man and Jersey), and 4 paging networks, while some country have up to a couple of dozen of each (China has more than 20 GSM networks). The Network Definition Language (NDL) has been designed to easily define and maintain routing information for these networks. Each sub-networks is described using a number of keywords and properties as shown on Figure 30.

```
define network <network name> {
        country name: <name>
        country code: <code>
        add prefix {
                <prefix list>
        }
        <override default routing>
        <force routing: network>
}
```

Figure 30: Network Definition Language syntax

A `define network` statement followed by a network name starts every definition. A typical network definition will include a prefix list, and some optional parameters. The prefix list, enclosed within an `add prefix { }` statement, contains all the sub-network addresses allocated for the network. The sub-network addresses are allocated by OFTEL in the UK and are available online for consultation. Warwick University also hosts a post-processed HTML version of the list suitable for computer processing. Other online sources such as the GSM World website and the ART website provide a list of country codes and routing information for the French networks. An import tool has been implemented to download and parse the documents and extract the relevant information, which is recreated automatically every month.

Wildcards are supported and used to specify a range of addresses, for example 4479760-6 means all addresses between and including 4479760 and 4479766. Figure 31 shows the whole range of addresses allocated to the UK GSM network Vodafone. At the time of writing the routing tables hold around 1500 different sub-network addresses split between 147 GSM and paging networks. Most networks offer a variety of submissions or access methods, be it by the hardware interface used or the submission protocol itself. Using Vodafone as an example again, the user has the choice between 9 access methods: 2 using modem dial-up, 3 using an X25 radio PAD, 3 using an X25 Kilostream and 2 using GSM mobile originated submission.
Each of these methods is handled by one or more NACP drivers, attached to a message queue. Each message queue is assigned a unique name which is used to control the routing for one or more networks. The submission methods, corresponding message queue and NACP driver are defined in a separate configuration file. Figure 32 depicts the various parameters available to configure a NACP driver:

```
define network Vodafone {
        country name: United Kingdom
        country code: 44
        add prefix {
                443700,443701,443702-9,443740-2,
                443744,443746-9,44378,443850-1,443852-3,
                443855,443857-9,44421,444481,444670-9,
                44468,444980-9,4477210-9,4477330-1,4477332,
                4477333-9,4477410-9,4477470-9,4477600-9,
                447767-8,4477750-9,4478180-9,
                4478670-9,4478790-9,4478870-9,4478990,
                4477690-9,4477700-9,447771,4477740-9,
                4477760-9,447778,4477800-9,4477810-9,447785,
                4477870-9,4477880-9,447798,4477990-9,
                4478310-9,4478330-9,4478360-9,
                4478800-9,4478810-9,4478840-9,
                4478991,4478992,4478993,4478994,4478995,
                4478996,4478997,4478998,4478999,4479010-9,
                4479090-9,4479790,4479791,4479792-3,4479794,
                4479795,4479796,4479797,4479798-9,4479900-9,
                448310-8,448362-3,448365-7
        }
}
```

Figure 31: sub-network addresses example for Vodafone UK

```
"<queue ID>" =>{
                start           => "<NACP driver start command>",
                stop            => "<NACP driver stop command>",
                restart         => "<NACP driver restart command>",
                kill            => "<NACP driver kill command>",
                description     => "<network name, extra information>",
                local_address   => "<NUA, MISDN, etc.>",
                charge_alias    => "<sender user name for billing purposes>",
                incoming path   => "<path for incoming messages and status reports>,
                incoming domain=> "<assign domain for incoming messages>",
                spoolformat     => "<spoolformat identifier>",
                maxchars        => "<max number of characters per message>",
                rules           => "<recipient pre-porcessing rules to apply>",
                options         => "<NoLinkChecks NoLocalAddress NoQueueChecks>",
},
```

Figure 32: Message queue definition

The parameters are defined as follows:

- start, stop, restart and kill: define the command to use to control the NACP driver;
- description: is a textual representation describing the queue and the driver;
- local_address: is the X25 NUA or MISDN address identifying the NACP driver;
- charge_alias: is for billing of incoming messages;
- incoming_path: is the path and method to use to pass on incoming messages;

- incoming_domain: is the email domain to use for incoming messages;
- options: is used to disable specific checks for that queue in the watchdog agent;
- maxchars: defines the maximum length of a single message. While short messages can be at most 160 characters, the length varies greatly amongst paging networks.
- spoolformat: is one of 5 file formats supported by the gateway. Most are now obsolete or have been superseded and are kept for backward compatibility only. The latest evolution is described further in this section.
- rules: while most networks access require the mobile numbers to be in international format (e.g. starting with the country code), a few legacy system are still in place (mainly for paging networks) that expect the number to be stripped of the country code and/or substitute the network part by one or more digits. For example: Vodapage access using TAP requires numbers starting by 441459 to start with 0 while number starting with 441460 should start with 1. The replace statement defines one or more regular expressions separated by ':'. The correct rules for the previous example would be: *rules: "/^441459/0/:/^441460/1/"*.

Figure 33 describes the configuration of a message queue for the Vodafone network using SMS2000 over X25 as the submission method. Note that some of the parameters are optional and default values are assumed if they are not specified.

```
"kstream3" =>{
               start           => "/e3/bin/sms2000d -k -b19200 -q kstream3 -d",
               stop            => "kill -TERM `cat /e3/log/pid/kstream3`",
               restart         => "kill -HUP  `cat /e3/log/pid/kstream3`",
               description     => "Vodafone, Mobile Alert NUA 23533420009401",
               local_address   => "23533420009401",
               charge_alias    => "www-sms",
},
```

Figure 33: Message queue configuration for Vodafone and SMS2000

The NDL also defines 2 additional statements to cope with routing of foreign networks and definition of static routes. Routing short messages to foreign networks relies on a roaming agreement being in place between the local network used for submission and the destination network. It also depends on the overall network architecture, and particularly the international exchange, to be configured properly (see section 3.3 for more details). While most of the time messages get through to their final destination, some networks require some adjustments. It might be necessary sometimes to force the routing of a network through another route than the default one (this would the case for example if the default network does not have a roaming agreement with the destination network). The `force routing <network name>;` statement is used in this case. Figure 34, depicts the configuration to route Digifone, an Eire based GSM network to be routed through Cellnet. The worst case scenario happens when a network simply blocks messages if they are not submitted

```
define network Digifone {
        country name: Eire
        country code: 353
        force routing Cellnet;
        add prefix {
                35386,35363
        }
}
```

Figure 34: forcing routing for a specific network

using one of its access methods[1]. The default route can not be used for submission in this case. The typical solution is to use a SIM from the foreign network (Itineris in this example), roaming on a UK based network (or ideally get a fixed link directly to the networks in question). The default routing will need to be overridden in order to define the access method for the network in question as seen on Figure 35.

```
define network Itineris_33 {
        country name: France
        country code: 33
        override default routing
        add prefix {
        33607,33608,33630,33654,33670,33671,
        33672,33673,33674,33675,33676,33677,
        33678,33679,33680,33681,33682,33683,
        33684,33685,33686,33687,33688,33689
        }
```

Figure 35: routing definition for foreign networks

6.3.5.2. Routing tables definition

Figure 36 describes a typical configuration for a gateway with full connectivity to all the UK based GSM and paging networks. For load sharing purposes, multiples queues can be specified and are separated by commas. Backup routes can also be defined, and must be separated by a '|'. At creation, the routing tables are initialised with the primary route for a network definition (i.e. kstream0-6 for Vodafone). Any backup routes, if defined, are used for resilience: the routing table are modified dynamically by the watchdog agent to cope with hardware failure, SMSC outage, etc.

```
Vodafone: (kstream0,kstream1)|(gsm0)
Cellnet: (kstream2)|(gsm1)
Orange: (kstream3)|(gsm2)|(modem0)
One2One: (kstream4)|(gsm3)|(modem0)
Vodapage: (kstream5)|(gsm0)
Vodazap: (modem0)
BT-Paging: (kstream6)
PageOne: (rpad0)

Default network: Vodafone
```

Figure 36: message routing definition

6.3.5.3. Static routing tables definition

On some occasions, it might be necessary to route message according to different criteria, other than the MISDN number of the recipient. This proves to be too restrictive for a number of applications such as:

- Reverse charging (e.g. recipient pays for reception of the message);
- Account, user or number based routing;
- Destination network based routing;
- Priority routing;
- Least cost routing.

The algorithm is based on overriding the MISDN based routing in favour of an key based routing. The key can be anything such as an account name, a user name, or a recipient MISDN. Additional second stage criteria can be used to define routes depending on the destination network or a reverse charge tariff. An example definition is depicted on Figure 37.

[1] The French GSM networks (SFR, Itineris and Bouygues Telecom) are a classical example.

```
"<key>" =>      {
                        "<network a>"   =>      {
                                                "<charge band 0>" => "<target>",
                                                "<charge band 1>" => "<target>",
                                                "<charge band 2>" => "<target>",
                                                "default" => "REJECT",
                                                "override_static" => 1,
                                                }
                        "<network b>"   =>      {
                                                "<charge band 0>" => "<target>",
                                                "default" => "REJECT",
                                                }
                        "<network c>"   =>      {
                                                "default" => "REJECT",
                                                }
                        "default"       =>      {
                                                "<charge band 0>" => "<target>",
                                                "default" => "REJECT",
                                                },
                },
```

Figure 37: static routing definition

The end target is either a message queue or a REJECT or CANCEL value. The REJECT target can be used to deny access to specific networks for certain accounts or users. It can also be used when all charge bands are not available across all networks. For example, network A offers band 1 to 4 except 3 while network B also offers band 1 to 6. In this case the static route can be defined to REJECT messages for band 3 and 5 and above on network A and for band 7 and above on network B. The CANCEL target simply uses the result from the normal MISDN based routing as a return target, in effect bypassing the static route. An example configuration is depicted on Figure 38.

```
"UoS" =>        {
                        "Vodafone"      =>      {
                                                "0" => "kstream0",
                                                "1" => "kstream1",
                                                "2" => "kstream2",
                                                "3" => "REJECT",
                                                "4" => "kstream3",
                                                "default" => "REJECT",
                                                }
                        "Cellnet"       =>      {
                                                "0" => "www1",
                                                "1" => "www2",
                                                "2" => "www3",
                                                "3" => "www4",
                                                "4" => "www5",
                                                "5" => "www6",
                                                "6" => "www7",
                                                "default" => "REJECT",
                                                }
                        "default"       =>      {
                                                "0" => "CANCEL",
                                                "default" => "REJECT",
                                                },
                },
```

Figure 38: using the REJECT target

Reverse charged band are defined from band 0 (e.g. not reversed charged) to band 10. Each band corresponds to a particular tariff ranging from 10p to £1. All messages having a destination bands other than 0 have to be spooled in the corresponding message queue; even if the NACP driver attached to the queue is reporting a link down. This ensures that the correct tariff is applied and the right person billed for the message submission.

Static routing for band 0 messages also takes into account the state of the link for the target message queue. To ensure minimum delay, the static route is cancelled if the link is down (this behaviour can be overridden using the override_static option).

The static routing algorithm is implemented as follows:

- For each <key> search for an entry in the static routes tables;
- If no entry is found then proceed with normal routing;
- If an entry is found then search for the destination network;
- If the destination network is not found then use the default entry;
- If no charge band have been specified then return the target for band 0
- Otherwise reject the message (no reverse route available for the tariff specified)

The implementation is very flexible and allows for some complex routing scenario.

6.3.5.4. *Message routing algorithm*

The complete routing algorithm is described on Figure 39.

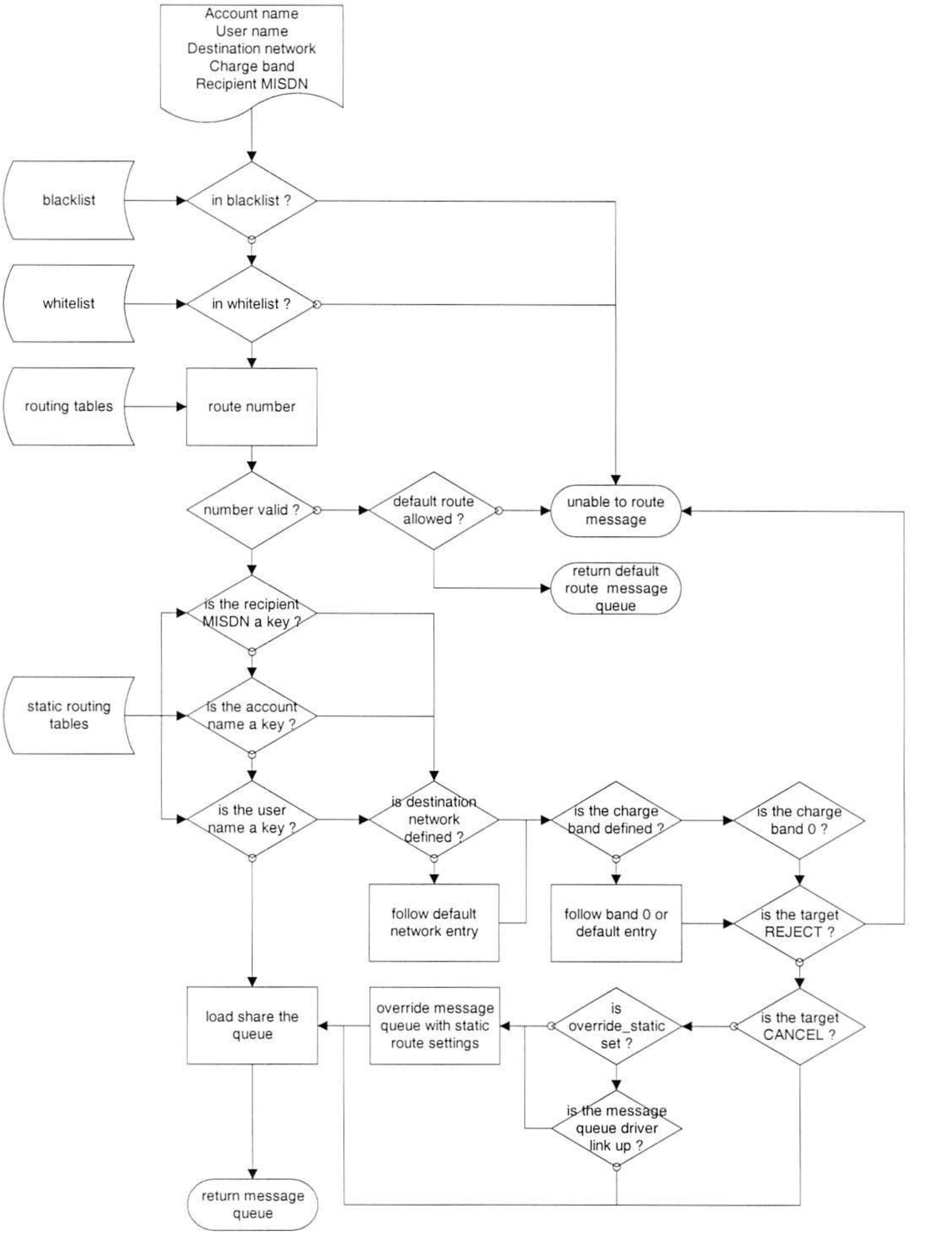


Figure 39: message routing algorithm

6.3.6. Message spooling

Upon routing, the messages are spooled to a queue according to the routing result, awaiting submission. There is one queue for each of the hardware interfaces connected to the gateway, with one independent submission process per queue. The messages queued are stored on the server's file system in a directory corresponding to the submission's interface name as depicted on Figure 40. Each directory contains a sequence file and a FIFO used as the inter-process communication mechanism between the MTP and the NACP. The MTP uses the sequence file to create a unique filename for each message spooled. As multiple instances of the MTP can be running simultaneously, access to the sequence file must be secured while reading and updating it. A kernel lock is requested before opening the sequence file. The sequence file typically contains a 6 digits number that is incremented by the MTP after every access. The kernel lock is then released and the message spooled to the FIFO and a backup file created in the current directory. The backup file is mainly a security measure to cope with failure of the NACP (e.g. driver crash). The spool process is depicted on Figure 41. The information contained in a spooled message is kept to a minimum with close attention paid to ease of access for the NACP. Each message spooled in the queue contains a series of information fields separated by a carriage return acting as an end of record. The message spooled carries information about the originator of the message, the message content and encoding and unique identifier as well as submission options - such as validity period and time to live in the system.

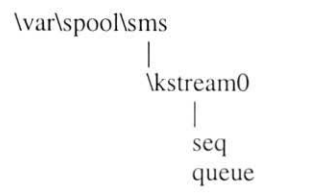


Figure 40: message queues representation on the file system

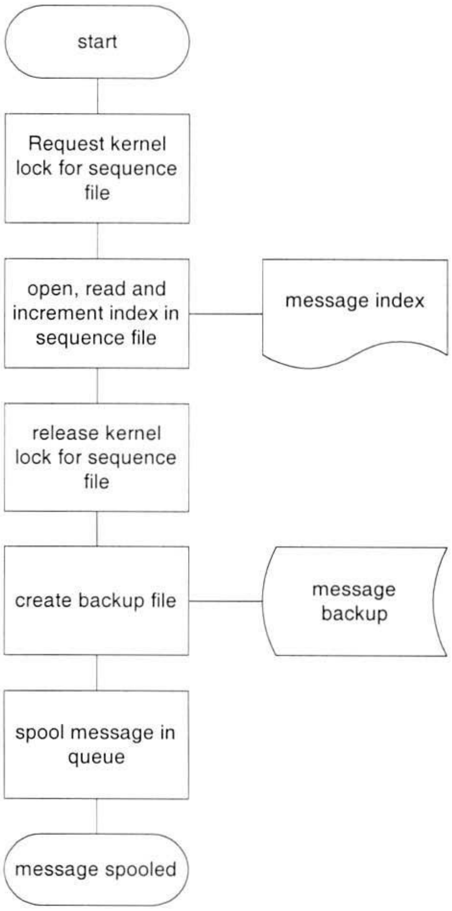

start
Request kernel lock for sequence file
open, read and increment index in sequence file
message index
release kernel lock for sequence file
create backup file
message backup
spool message in queue
message spooled

Figure 41: Message spooling process flow

The following table describes the information spooled per message in the queue.

Field	Description
Number	MISDN or pager number of the recipient
Message	Hex encoded message
Reply path	Reply path for confirmation of submission and delivery (if applicable), formatted as path type and originator information. Valid paths include: • SMTP • HTTP/S • SMPP • TAP • XML • SMPP
Originating address	Used to set the originating address (e.g. sender) of the message when supported by the destination network. This enables user to send messages with an alphanumeric (i.e. company name, MISDN number) originating address. This is used to brand messages or as session information for 2 way applications
Identifier	Message identifier, composed of: 1. Sequence index 2. Server identifier 3. Interface identifier 4. Epoch The sequence index is a 6 digits number extracted from the sequence file introduced earlier. The server identifier is the local hostname hex encoded (i.e. colossus), so is the interface identifier which relates to the interface name (i.e. kstream0). Finally the last field is the time the message was spooled in epoch format (in seconds since the 1st of January 1970).
User key	This allows the end user to tag a particular transaction and match a confirmation of submission or delivery with the original message.
Time to live	Absolute validity period in epoch format. The watchdog agent expires message when the time to live is exceeded.
Validity period	Absolute time to live in the system in epoch format. This field is used by the watchdog agent to expire messages still not submitted after a set time. This usually indicates a formatting problem with the message and the incapacity of the driver to submit it.
Submission attempt	Spool attempt, the field is incremented every time the message is spooled. This field is used by the watchdog agent to dynamically modify routing and detect failed drivers.

Options	**Option field. 8-bit array defined as follows:** • **Bit1: confirm submission if set** • **Bit 2: confirm delivery if set** • **Bit 3: set User Data Header indicator in the short message PDU for submission** • **Bit 4: statically routed message, do not respool** • **All other bits reserved**
Data coding scheme	**Data coding scheme as of ETSI 03.40 (default 00). This field is typically used to submit binary content such as ringtones or WAP WDP packets.**
Protocol identifier	**Protocol identifier as of ETSI 03.38 (default 00). This field is typically used in conjunction with the Data Coding Scheme for advanced short messaging with binary content or to set a message class (to replace a previously sent message for example).**
Media type	**8-bit array identifying the type of message:** **0000 0000 : plain text** **0000 0001 : RT / ringtone** **0000 0010 : OL / operator logo** **0000 0011 : SL / service loading** **0000 0100 : AD / VCard** **0000 0101 : AP / VCal** **0000 0110 : BM / Bookmarks** **0000 0111 : IA / Internet Access Point** **0000 1000 : EC / EmaiL configuration** **0000 1001 : PM / Picture Messaging** **0000 1010 : GL / Group Logo** **0000 1011 : SS / ScreenSaver** **0000 1100 : AS / AnimatedScreensaver** **0000 0110 : reserved** **to** **1111 1111 : reserved**
Authentication details	**Needed authenticate messages when queuing from one server to another**
Service centre	**For mobile originated submission, change the default service centre (e.g. SMSC) to use.**
Service centre type	**Sets the SMSC address type**

6.4. The Network Access Control Part

The NACP implements all the necessary network drivers (i.e. TAP, SMS2000, ETSI0705, etc.) to communicate with the SMSC using the communication devices. At this stage the original content has been formatted using the user profile and all the necessary options for delivery have been set. The NACP is constantly monitoring the state of the FIFO file it is attached to for more data to be read. Once the event has been triggered a new message is extracted from the queue; the content of the message is then encoded according to the submission protocol used and awaits submission. The NACP has been designed with reusability and flexibility in mind. While this has some quite heavy implications in terms of initial time to production, the benefits are then quite clear when a new protocol has to be implemented or a new IPC method tested between the MTP and the NACP. The next two sections will give an overview of the building blocks of the NACP, and describe the finite state machines which form the basis of all the drivers.

6.4.1. Protocol independent design

The drivers have been designed with as few specific parts as possible; while the protocol specific state machine is inevitable, it is very easy to create a design where all components are so inter-dependent that reusability is very difficult if not impossible. Figure 42 depicts the building blocks of all the drivers. Most blocks perform independent tasks and could easily be implemented as separate threads – the added complexity however, might not offer any significant benefits to justify it – an outgoing message travels through the layers from top to bottom triggering events in the other blocks. Events will in turn update submission statistics, log submission information, etc. The message will enter the driver state machine only when the current state allows submission (typically the driver is connected to the SMSC and ready to send). Every primitive used by the driver to modify its state, submit a message, and so on, is protocol specific and implemented in a separate state machine. This part and this part only needs changing when implementing a new protocol. Incoming messages are received asynchronously from the SMSC and passed as raw data to the incoming message agent for further processing. Once again a file or database store is used for that purpose.

The MTP stores formatted message in a queue (physically a directory on the local file system) and triggers an event in the NACP using one of the mechanism implemented – either a FIFO or a UNIX signal.

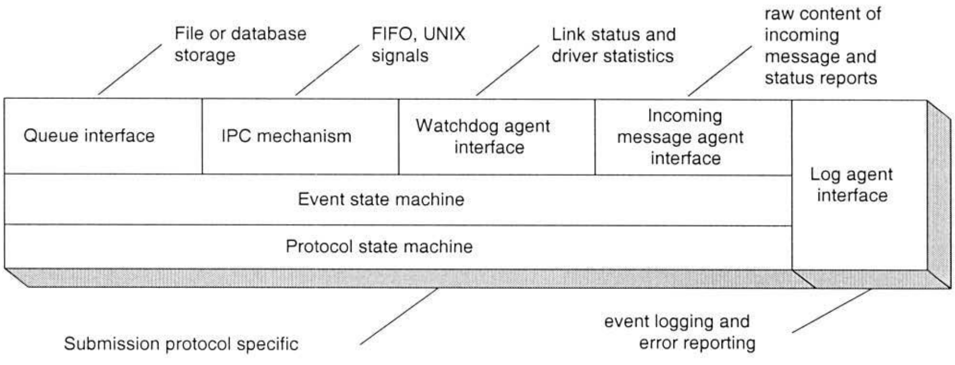

File or database storage
FIFO, UNIX signals
Link status and driver statistics
raw content of incoming message and status reports
Queue interface
IPC mechanism
Watchdog agent interface
Incoming message agent interface
Log agent interface
Event state machine
Protocol state machine
Submission protocol specific
event logging and error reporting

Figure 42: NACP building blocks

The NACP also includes an interface to the watchdog agent and exports various statistics such as current state, number of message sent, received or failed, link status etc. The link status relates to the signal quality and BER rate of a wireless link or the last time a successful connection to the SMSC was achieved for a leased line.

The log agent interface simply updates a system wide log file with occurrences of events in the NACP such as message submission or delivery.

6.4.2. The NACP state machines

The basic idea is that only a few states are necessary for any short message submission protocols to be implemented. The driver will always have one of these states, while transition to another state is ruled by one or more conditions. The state machine is implemented as a two dimension array of pointers to functions, similar to many used in communications protocols such as TFTP.

The specifics of each protocol and the hardware used (X25 leased line, V34 serial connection) are implemented in the bottom layer of the NACP. The primitive used to connect to the SMSC, submit a message or hang-up vary greatly between protocols due ton the lack of standardisation. An XML based API to front-end the SMSC will surely be a welcome addition to each network operator offering. The driver state machine calls primitives in the protocol state machine for every transition between states. Figure 43 describes the driver state machine and the possible transitions between them. Every transition can be associated with a primitive call and a condition for the transition to happen. Changing from the offline to the online state require a call to the connect() subroutine and for the SMSC to acknowledge the request. If one or the other fail the driver will simply stay in the same state and retry and eventually transit to another state.

Six states have been defined and are used in the driver state machine:

1. Init
2. Offline
3. Online
4. Sending
5. Receiving
6. Shutdown

Init and Shutdown ane the entry and exit point of the state machine when the driver is started and stopped (usually on reception of a SIGTERM signal). The driver stays in the init state for as long as it takes to initialise the hardware parameters. A GSM modem for example is queried for the number of slots in the SIM card, the default SMSC address to use, while a TCP/IP connection needs the SMSC name to be resolved to an IP address. Once all the parameters are initialised, the state changes to offline and the driver is ready to process queries.

The shutdown procedure is simple but strict, some protocols such as SMPP require that a command is sent to the SMSC prior to a disconnection, this usually avoid having stale processes on the SMSC side and waste computing resources. Further connections might also fail if the appropriate disconnection procedure is not followed. Consequently the shutdown state is only reachable from 2 states: init and offline, when the driver is not connected to the SMSC. Should the driver be in any of the other states it would have to satisfy the transitions to the offline or init state before being allowed to shutdown (see Figure 44 for details).

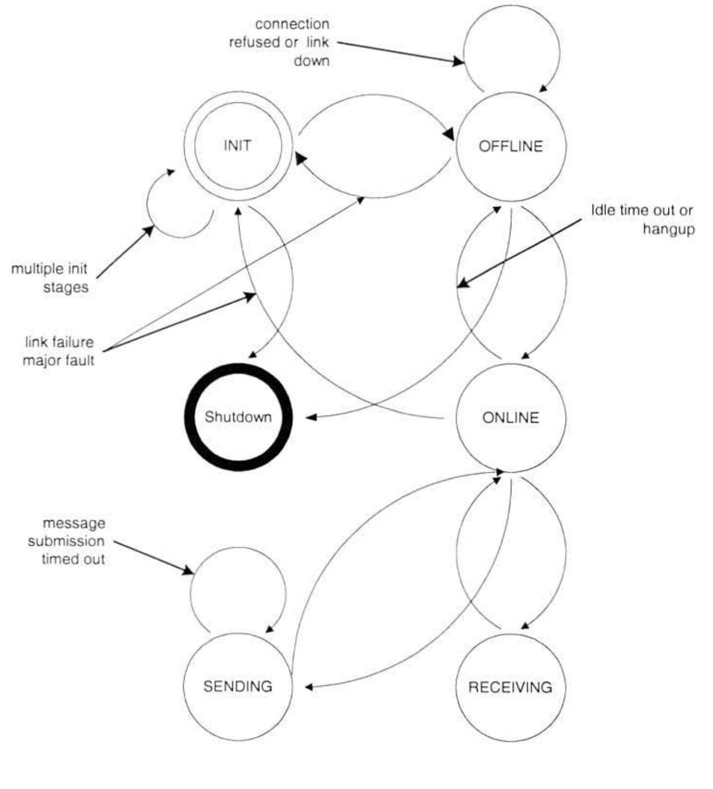


Figure 43: NACP driver state machine

Shutdown scenario:

- The driver is busy sending a burst of messages, the current state is sending;
- A request for shutdown is received;
- The driver queues the request and changes its target state to shutdown;
- The last message submitted is acknowledged, the current state changes to online;
- The state machine is accessed with a current state of online and a target state of shutdown. As a result the disconnect() call is made;
- The SMSC acknowledges the request for disconnection the current state is now offline;
- The state machine is accessed with a current state of offline and a target state of shutdown, the shutdown() call is made;
- The driver logs an entry and exits.

Offline is the idle state, when the driver is not connected to the SMSC and there are no pending messages. The driver usually stays in that state for a period of time then initiate a connection to the SMSC to check that the link is still up (e.g. SMSC reachable and connection request acknowledged). The SMPP submission protocol for example requires the driver to be constantly bound (e.g. connected) to the SMSC and therefore has built in primitives to enquire the status of the link. In this last case online becomes the idle state.

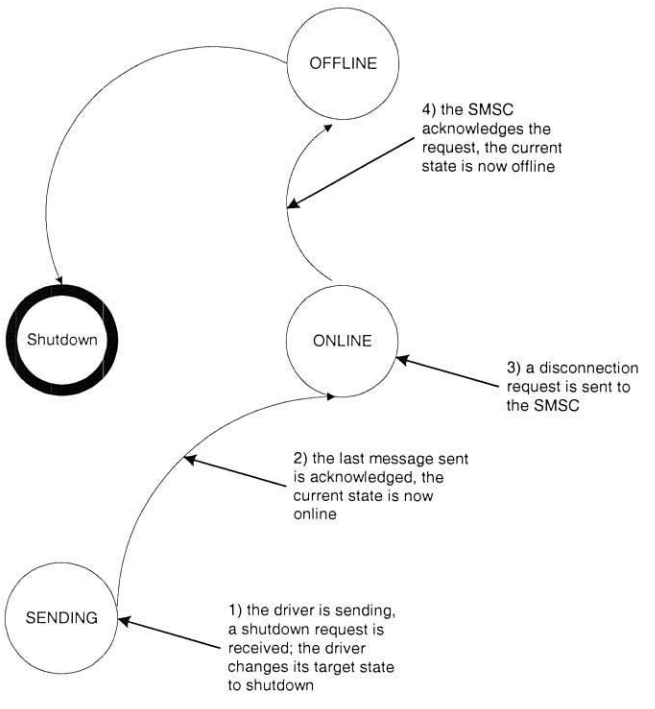

OFFLINE
4) the SMSC acknowledges the request, the current state is now offline
Shutdown
ONLINE
3) a disconnection request is sent to the SMSC
2) the last message sent is acknowledged, the current state is now online
SENDING
1) the driver is sending, a shutdown request is received; the driver changes its target state to shutdown

Figure 44: NACP shutdown scenario

The driver is considered online when a successful connection is made to the SMSC. This is achieved in a number of ways depending on the hardware and protocol used. It can take from a few milliseconds (i.e. using leased lines) to a few seconds (i.e. using dial-up modems). Once the connection is established, the transition to the next state depends on the event that triggered the change: the driver will stay online for a period of time then hang-up or process any ending request for message submission. One or more requests for submission can be queued asynchronously at any single time. If one or more messages are awaiting submission the driver will change its target state to sending and enters the protocol state machine.

The sending state is reached once a successful connection to the SMSC is achieved and an outgoing message is pending in the queue. The protocol state machine is accessed at this stage to read the message from the queue, encode the content according to the submission protocol used and finally submit the appropriate primitives to the SMSC. Conversely, the receiving state is reached when an incoming message primitive is received from the SMSC. The driver enters the protocol state machine to acknowledge the message, extract the content and pass it to the incoming message agent for further processing.

6.4.3. Character set translation

An important issue when designing a gateway architecture between applications using different character sets is the corresponding mapping. This is essential to prevent message corruption, as different kinds of computer systems may internally represent the same character in a completely different way. Therefore, the gateway translates every character in the original message into its corresponding value in the GSM default [ETSI 3.38 1996]. We describe the solution implemented in our design to provide a mapping between the ISO8859-1 and the default SMS character set. This mechanism can easily be extended to cover any character set mapping.

The SMS defines its own default character set [ETSI 3.38 1996] but leaves to the GSM network operators the ability to specify the extended 8- or 16-bit character set. The default character set is 7-bit wide thus defining 128 characters, however some manufacturers implement 8- or even 16-bit character sets and integrate them to the design of their phones. Some mobile phones can support up to 7 different European Languages while mobile phones in the Far East can support Chinese and Thai. (Nokia was the first manufacturer to launch a mobile phone capable of displaying Far East characters, and their handset uses a 16-bit character set and has a bigger and wider display).

Table 12: ETSI-0338 default character set

	0	1	2	3	4	5	6	7
0	@	Δ	SP	0	¡	P	¿	p
1	£	_	!	1	A	Q	a	q
2	$	Φ	"	2	B	R	b	r
3	¥	Γ	#	3	C	S	c	s
4	È	Λ	¤	4	D	T	d	t
5	é	Ω	%	5	E	U	e	u
6	Ù	Π	&	6	F	V	f	v
7	ì	Ψ	'	7	G	W	g	w
8	ò	Σ	(	8	H	X	h	x
9	Ç	Θ	)	9	I	Y	i	y
A	LF	Ξ	*	:	J	Z	j	z
B	Ø	SP	+	;	K	Ä	k	ä
C	ø	Æ	,	<	L	Ö	l	ö
D	CR	æ	-	=	M	Ñ	m	ñ
E	Å	ß	.	>	N	Ü	n	ü
F	å	É	/	?	O	§	o	à

The default character set for SMS is shown in Table 12. This default alphabet is mandatory and must be supported by all MSs and SMSCs. The character set contains all the ASCII letters and numbers, some German, Greek and French accentuated characters, symbols and punctuation signs and three control characters (namely carriage return, line feed and space), all provided by the 7 bit ISO standard.

However, most computer systems use more flexible 8bit character sets including the ISO 8859 Latin alphabets 1 and 2 (Western and Eastern European languages). Multi-byte character sets such as Unicode or the Korean, Japanese or Chinese national character set are not addressed in this document.

There are currently ten different ISO 8859-x standards each supporting a particular group of languages. The Latin alphabets are 8-bit, 256 characters sets. The left half of the character set (i.e. first 128 characters) is the same as the ASCII characters set, while the right half contains language specific graphical characters.

The ISO8859-1 Latin 1 character set contains graphical characters used in at least 44 countries and can support languages such as Danish, Dutch, English, Finnish, French, etc. It has become increasingly popular in the electronic messaging world. The ISO 8859-1 Latin 1 character set is listed is shown in Table 13.

Table 13: The ISO8859-1 Latin 1 alphabet

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p				°	À	Đ	à	ð
1	SOH	DC1	!	1	A	Q	a	q			¡	±	Á	Ñ	á	ñ
2	STX	DC2	"	2	B	R	b	r			¢	²	Â	Ò	â	ò
3	ETX	DC3	#	3	C	S	c	s			£	³	Ã	Ó	ã	ó
4	EOT	DC4	$	4	D	T	d	t			¤	´	Ä	Ô	ä	ô
5	ENQ	NAK	%	5	E	U	e	u			¥	μ	Å	Õ	å	õ
6	ACK	SYN	&	6	F	V	f	v			¦	¶	Æ	Ö	æ	ö
7	BEL	ETB	'	7	G	W	g	w			§	·	Ç	×	ç	÷
8	BS	CAN	(	8	H	X	h	x			¨	¸	È	Ø	è	ø
9	HT	EM	)	9	I	Y	i	y			©	¹	É	Ù	é	ù
A	LF	SUB	*	:	J	Z	j	z			ª	º	Ê	Ú	ê	ú
B	VT	ESC	+	;	K	[	k	{			«	»	Ë	Û	ë	û
C	FF	FS	,	<	L	\	l	|			¬	¼	Ì	Ü	ì	ü
D	CR	GS	-	=	M	]	m	}			-	½	Í	Ý	í	ý
E	SO	RS	.	>	N	^	n	~			®	¾	Î	Þ	î	þ
F	SI	US	/	?	O	_	o	DEL			¯	¿	Ï	ß	ï	ÿ

We now describe the mapping mechanism implemented. To illustrate the procedure we will consider a user sending an electronic mail using the ISO8859-1 Latin 1 characters set to another user's handset only supporting the default SMS alphabet. Assume that User A sends a message to User B containing the accentuated character 'é' of hexadecimal value E9 in the Latin 1 character set (see Figure 45). Without any mapping procedure the message would get at best accepted (but corrupted) by the message centre and at worse purely rejected.

The mapping procedure uses a set of translation tables such as the one depicted on Table 14 and Table 15, to map the value of characters from one alphabet to another. In the case of the accentuated character 'é' the procedure converts its hexadecimal value from E9 (ISO8859-1 value) to 05 (ETSI0338 value). As the mapping procedure is translating from an 8-bit alphabet to a 7-bit alphabet some characters in the original message can not be translated. In this case the gateway tries to replace the original character by its closest look alike (i.e. U instead of Û), or if no acceptable compromise can be found the character space will be used instead (i.e. character ©).

Table 14 : Translation table from ISO8859-1 to ETSI-0338 (first half)

	0		1		2		3		4		5		6		7	
0	NUL		DLE		SP	SP	0	0	@	@	P	P	`	`	p	p
1	SOH		DC1		!	!	1	1	A	A	Q	Q	a	a	q	q
2	STX		DC2		"	"	2	2	B	B	R	R	b	b	r	r
3	ETX		DC3		#	#	3	3	C	C	S	S	c	c	s	s
4	EOT		DC4		$	$	4	4	D	D	T	T	d	d	t	t
5	ENQ		NAK		%	%	5	5	E	E	U	U	e	e	u	u
6	ACK		SYN		&	&	6	6	F	F	V	V	f	f	v	v
7	BEL		ETB		'	'	7	7	G	G	W	W	g	g	w	w
8	BS		CAN		(	(	8	8	H	H	X	X	h	h	x	x
9	HT		EM		)	)	9	9	I	I	Y	Y	i	i	y	y
A	LF	LF	SUB		*	*	:	:	J	J	Z	Z	j	j	z	z
B	VT		ESC		+	+	;	;	K	K	[	(	k	k	{	(
C	FF		FS		,	,	<	<	L	L	\	/	l	l	|	:
D	CR	CR	GS		-	-	=	=	M	M	]	)	m	m	}	)
E	SO		RS		.	.	>	>	N	N	^	SP	n	n	~	-
F	SI		US		/	/	?	?	O	O	_	_	o	o	DEL	

Table 15: Translation table from ISO8859-1 to ETSI-0338 (second half)

	8		9		A		B		C		D		E		F	
0						SP	°	o	À	A	Ð	D	à	à	ð	o
1					¡	¡	±	SP	Á	A	Ñ	Ñ	á	a	ñ	ñ
2					¢	c	²	2	Â	A	Ò	O	â	a	ò	o
3					£	£	³	3	Ã	A	Ó	O	ã	a	ó	o
4					¤	¤	´	'	Ä	Ä	Ô	O	ä	ä	ô	o
5					¥	¥	µ	u	Å	Å	Õ	O	å	a	õ	o
6					¦	:	¶	SP	Æ	Æ	Ö	Ö	æ	æ	ö	ö
7					§	§	·	·	Ç	Ç	×	x	ç	Ç	÷	/
8					¨	SP	¸	,	È	E	Ø	0	è	è	ø	Ø
9					©	SP	¹	1	É	É	Ù	U	é	é	ù	ù
A					ª	a	º	o	Ê	E	Ú	U	ê	e	ú	u
B					«	<	»	>	Ë	E	Û	U	ë	e	û	u
C					¬	-	¼	SP	Ì	I	Ü	Ü	ì	ì	ü	ü
D					-	-	½	SP	Í	I	Ý	Y	í	i	ý	y
E					®	SP	¾	SP	Î	I	Þ	SP	î	i	þ	SP
F					¯	-	¿	¿	Ï	I	ß	ß	ï	i	ÿ	y

▨	Indicates that an illegal character had to be discarded
	Indicates that an approximation had to be made

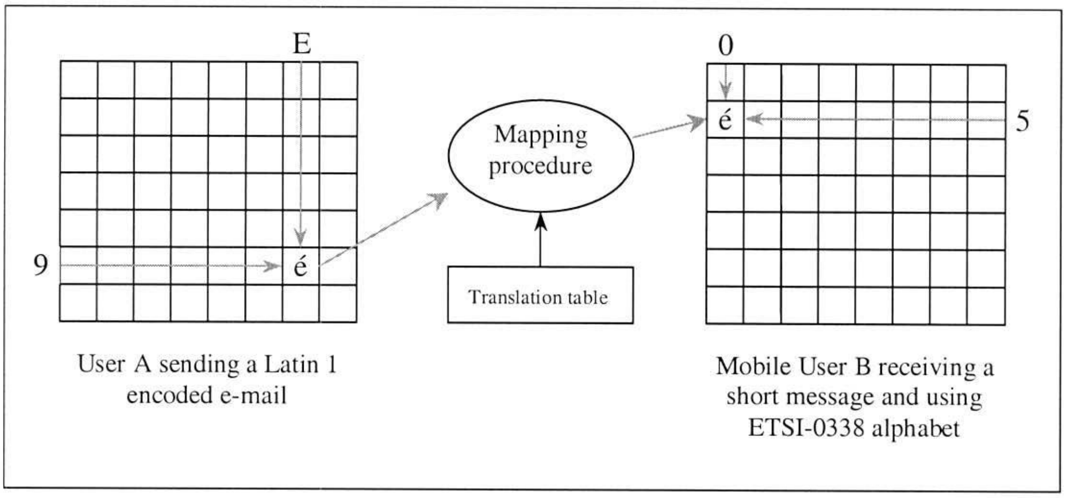

E
0
é
5
Mapping procedure
9
é
Translation table
User A sending a Latin 1 encoded e-mail
Mobile User B receiving a short message and using ETSI-0338 alphabet

Figure 45: character set mapping procedure

6.5. Transaction and event logging

Each step in the translation process is logged for statistical, invoicing or debugging purposes, by a separate process: the ***Logging agent***. A global file is used to log the transactions in the following format:

<date><*space*><time><*space*><entry>

where:

<space> ::= 'a blank space character'
<semi column> ::= 'a semi column character ;'
<entry> ::= <entry type><semi column><entry data>
<entry data> ::= ' a semi column separated value list'

The date and time format is inspired from a standard date format, but the offset from GMT is not included (all time are assumed to include BST alterations) and the date fields are in the following order: day-month-year. Moreover, the resolution of the time is down to a 1-microsecond granularity to allow more accurate timing.

Date and time example: ***17-07-1999 13:47:56:678765***

Entry type	Description
C	**Information related to the processing in the MTP**
S	**Information related to processing in the NACP**
D	**Information related to message delivery to the GSM terminal**
F	**Information related to message submission failure**
I	**Information related to incoming short message**
R	**Information related to message re-spooling and submission retries**

The entry data varies depending on the entry type and is detailed below:

C	ID	Interface ID	Account	Originator address	Recipient address	Start of conversion time stamp	End of conversion timestamp	Submission interface	Submission protocol

S	ID	Start of submission time stamp	End of submission timestamp	Encoded message

D	ID	Short message delivery time stamp

F	ID	Short message submission failure timestamp	Error cause

R	ID	Previous ID	Short message resubmission timestamp	Submission interface	Submission protocol

I	ID	Short message reception timestamp	Originator address	Encoded message

The fields are defined as follow:

E3 ID	**Unique Id assigned to a message for its whole life cycle in the system.** **Format as follows:** **<sequence number>-<interface>-<epoch>** **epoch being the number of seconds elapsed since 1st of January 1970**
SMTP ID	Unique Id assigned to an e-mail by the gateway's mail server. Format as follows: <message id>@<SMTP server fully qualified name or IP address>
Account	Unique identifier, later used for invoicing purposes
Originator address	Depending on the entry type, it will be an e-mail address or a GSM terminal ID
Recipient address	The recipient's GSM terminal ID
Submission interface	Identifies the interface used for submission (i.e. GSM0, Kstream0, etc.)
Submission protocol	Any of the following: • SMS2000 • 0705 (block or PDU mode) • SMPP • TAP • CIMD • Etc...
Encoded Message	Hex encoded message (the field is optional)
Service ID	gateway services ID.
Error cause	Detailed error cause: • Delivery expired • Terminal unavailable • Connection rejected • Etc...

6.6. Log example

A typical transaction on the gateway would generate the following output in the log file:

```
04-07-1999 13:45:01:823454 C;039844-kstream1-932536254;SMTPID;sms@envelos.com;
pageadmin;04-07-1999 13:45:00:037865;04-07-1999 13:45:01:627563:kstream1:SMS2000
.
.
.
04-07-1999  13:45:03:946654  S:  039844-kstream1-932536254;04-07-1999  13:45:01:945364;04-07-1999
13:45:03:324654;6465656565664636
.
.
.
04-07-1999 13:45:08:345454 D:039844-kstream1-932536254;04-07-1999 13:45:07
```

Note that the time stamp for the delivery of the short message has only a resolution of 1 second. This timestamp is extracted from the notification of delivery message returned by the SMSC upon successful delivery of the message to the GSM terminal.

A successful e-mail to short message conversion will result in the following sequence of entry in the log file: C, S, and D (if requested). Should the message submission fail and the gateway had to retry or use a different interface, the sequence will then be C, S, (F, R)^n, S, and D (if requested).

6.7. Alerting and monitoring

The ***Watchdog agent*** monitors the gateway and periodically performs a range of tests to check for resources problems and identify NACP driver problems and modify the routing dynamically. The agent will typically perform the following tasks:

- Report problems by short messages and email to the administrator(s);
- Check the status of the NACP drivers and the links to the SMSC's;
- Check the amount of free memory left;
- Check the number of inodes available;
- Check the number of file descriptors available;
- Check the total number of processes running;
- Check the load average on the server;
- Check on the available disk space;
- Check that the SMTP and HTTP interface are running;
- Check the health of the raid array;
- Recreate the gateway's user database files when out of date;
- Recreate the routing tables if the gateway configuration changes;
- Clean out expired messages from the NACP queues;
- Re-spool messages that meet certain criteria;
- Warn if any queues are building up in size;
- Warn if any queues are not moving;
- To dynamically take queues in or out of the current routing depending on status or the load;
- To switch the current routing for a network when one or more NACP drivers are reporting faults and /or outages.

6.8. Conclusion

This chapter has presented a gateway architecture designed and implemented to provide connectivity between short messaging and paging services and electronic mail. The integration of mobile messaging with the Simple Mail Transfer Protocol (SMTP) allows email users to send messages to mobile terminals using standard email packages. The architecture provides a consistent gateway between email and mobile messaging services by providing mapping of electronic mail addresses and mobile terminal numbers. The gateway allows seamless delivery of messages across different cellular networks. It also implements novel features often overlooked by similar products: It provides robust messaging for intranet or mission critical applications and is suitable when high message throughput or fail-safe reliability is required. The main issues related to the message translation process have been described, such as character set mapping and electronic message format. The protocol architecture of the gateway has also been detailed.

7. Experimental evaluation

7.1. Introduction

This section presents the methodology and results of a study conducted to characterise the performance of the e-mail to SMS gateway introduced in the previous chapters. The study is based on a simple test scenario with the gateway in a basic configuration (see **Error! Reference source not found.**).

The analysis is based on benchmarking using the logging information provided by the gateway via the log agent and a protocol analyser (frame capture); and will also attempt to develop a source model to compare theoretical and real performance. The complexity of the system and the number of components to be studied mean that developing a source model will be extremely difficult if not impossible without using measurement to quantify some of the component's properties: the performance of the gateway as a whole or of any of its components is likely to be highly related to the software (operating system, C compiler used, memory model, etc.) and hardware configuration (MISC, RISC, 32 or 64 bits processors, etc.). Moreover, measurement of the real system with all components in place would reveal unexpected interactions that may not be observed by studying the same components individually.

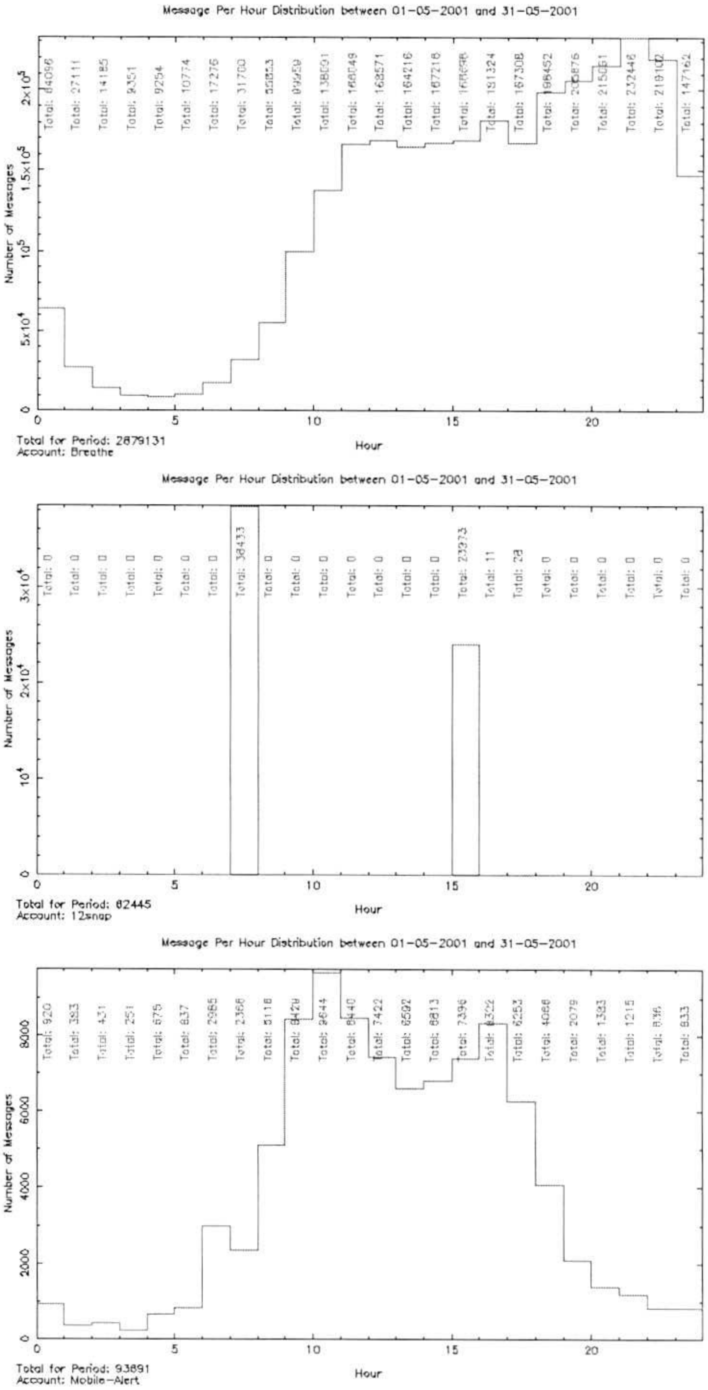
7.2. Methodology

The test bed configuration is rather unusual in the way it uses live gateways to generate emails pseudo-randomly. Admittedly this is a non standard approach as a packet generator programmed to follow a specific pattern - for distribution in time or inter-arrival rate – is more frequently used in this type of testing. However what the test is trying to demonstrate here is real life behaviour under the conditions that will appear during a typical day, week or month.

A Monday morning when all secretaries e-mail reminders of the week's meetings, will be busier than a Friday morning being the last day of the week. It is fairly easy to guess that some months will be busier than others, typically the July-August period being the quieter and the first few months of the year the busier, as people tend to work more in the spring-summer month than in the autumn-winter months leading to Christmas. Apart from the main holiday periods, it is fairly safe to assume that every week will be similar in terms of traffic generated and distribution of this traffic throughout the day.

The live gateways running as a four server cluster serve a vast number of customers and process high numbers of messages everyday. The customer profile and main activity of business will in most cases relate to a traffic distribution that can be predicted. Wireless marketing companies are likely to send bulk quantities of messages to a large recipient list resulting in a sudden burst of messages in the queues. An ISP offering free message for its customer base will generate a more steady flow of messages. Email alerting services - who forward copies of email as short messages – are more relevant to this experiment as the flow of traffic is generally concentrated within office hours and with similar volumes every day of the year.

The next figures illustrate by an example the message volumes generated hourly by three customers falling in the previously mentioned categories. The distribution of the submitted messages is calculated using the data stored by the live gateways. In essence, it represents the distribution of the departure rate of the NACP (or submission rate to the SMSC).



ISP model

Wireless marketing company model

Email alerting company model

Figure 46: hourly message submission distribution example

For the purpose of this experiment, I have decided to use the data generated by an email alerting company as the traffic to characterise the delays in each layer of the gateway. The distribution of their traffic was assessed over the space of a year - Sept 2000 to Aug 2001 - and found to be uniform enough to show a repeating pattern. Over the period the average monthly message volume was 91500, with a lower and upper boundary at 80200 and 108000 respectively (see Figure 47).

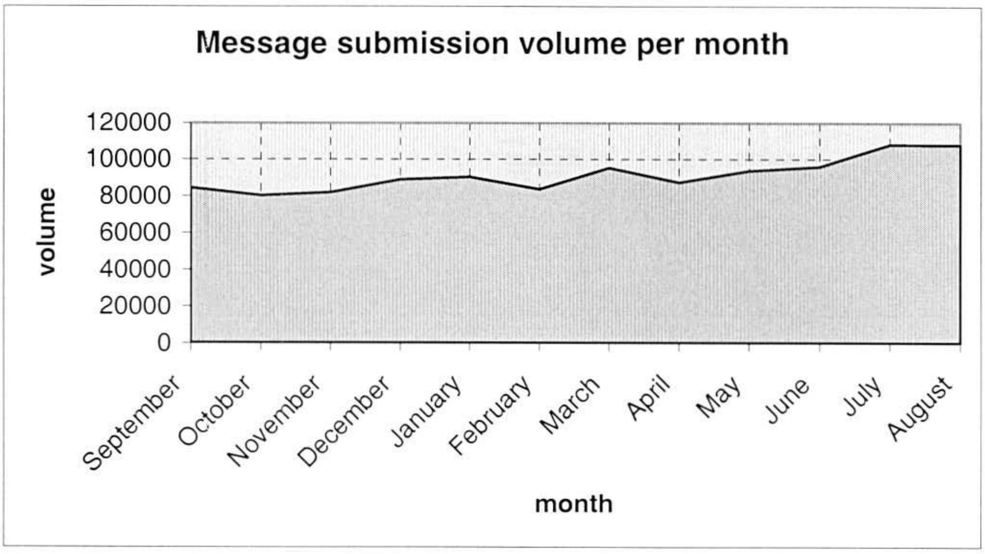


Figure 47: monthly message submission volume from September 2000 to August 2001

Figure 48 shows the detailed distribution of the volume of messages submitted by the customer over the same period. The reader will notice that the maximum number of messages sent within a specific month is used to scale the Y axis and as a consequence the scale varies from one graph to the other. What is more important however is that the distribution of the volume of messages throughout the months shows a recurring pattern. The daily volumes are also fairly constant with the expected drop at week-ends.

An analysis of the distribution of the message submission volume per hour is used to estimate the parameters for a distribution function. The analysis was carried out using the data recorded for the month of May 2001.Figure 49 shows the graphical representation of the distribution: most of the messages are submitted during office hours (bar the few late workers still sending at 9pm), with a peak around lunch time. The pattern seems to follow a classical normal distribution with a mean μ and a standard deviation σ.

The equation for the normal density function is given by:

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \quad e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The distribution is centred around its mean μ, at which it peaks. One can estimate that the mean of the measured distribution is reached at the time when the volume of messages submitted is equal to the number of messages left to submit for the day.

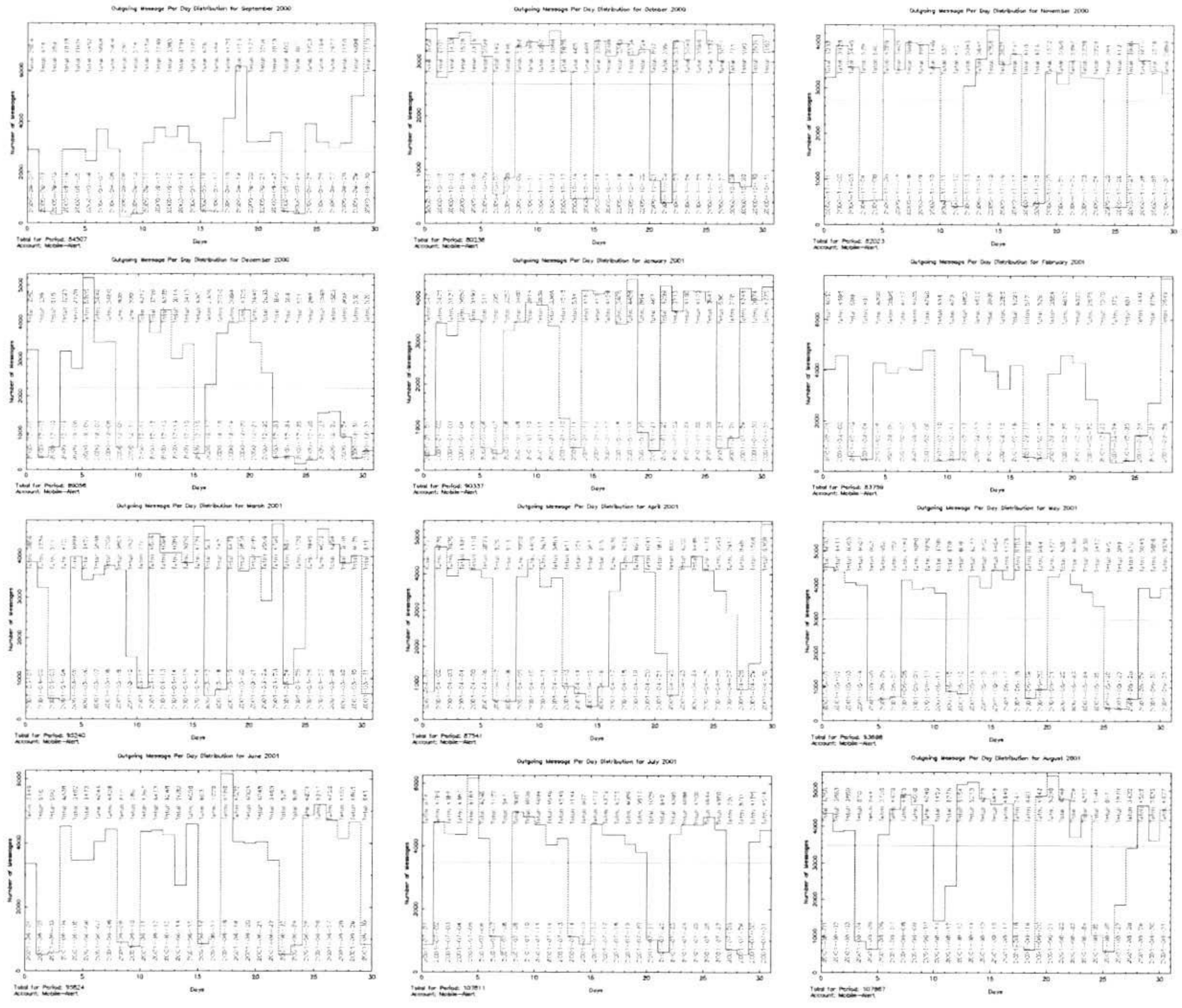


Figure 48: monthly message submission distribution example

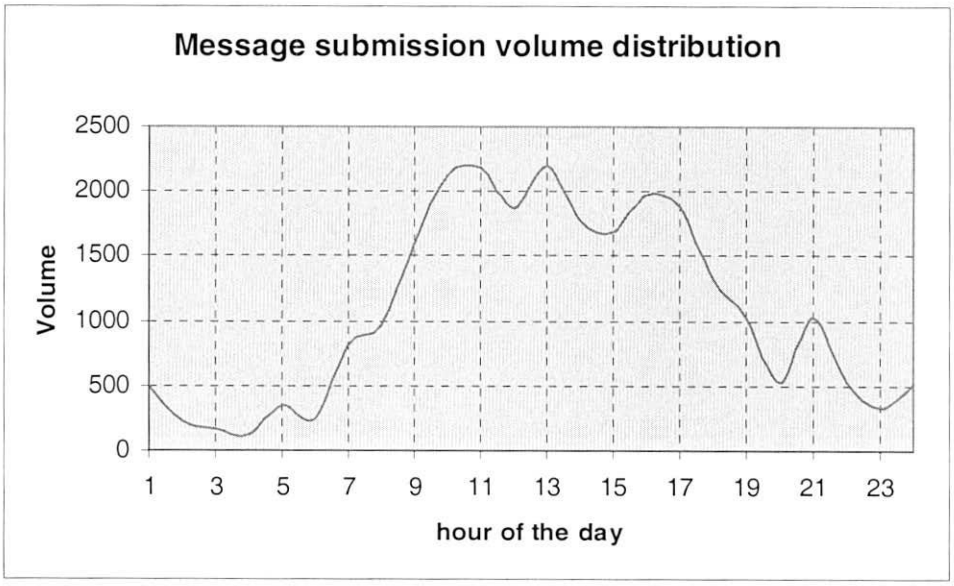


Figure 49: measured message volume distribution

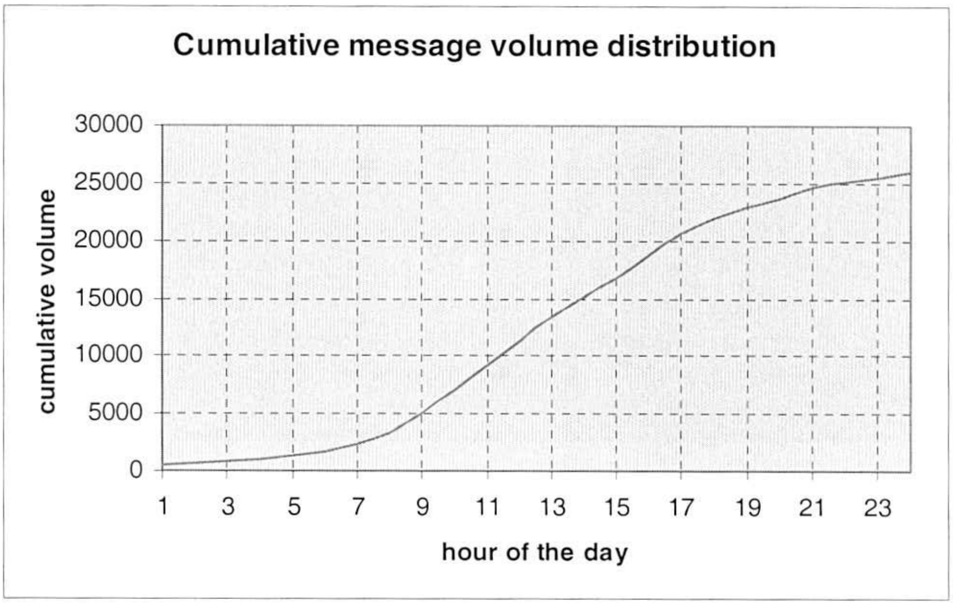


Figure 50: measured cumulative message volume distribution

The cumulative distribution of the measured message volume can be used to give a reasonable estimate of the mean. While the total number of messages sent over the period is 25950 the distribution function should be centred around the hour when half of that volume has been submitted (12975 messages).Figure 50 shows that the measured cumulative volume of 12975 is reached sometimes between the 12th and 13th hour, with the exact time given by:

$$\mu = H + \frac{\sum_{h=0}^{h=H} V_h - \sum_{h=0}^{h=H-1} V_h}{\sum_{h=0}^{h=H+1} V_h - \sum_{h=0}^{h=H-1} V_h}$$

where: μ is the mean of the distribution
T is the hour just before the mean is reached
V_h is the volume of messages submitted during the hour h

Hour h	Cumulative volume	Volume V_h	Hour h	Cumulative volume	Volume V_h
1	503	503	13	13350	2201
2	723	220	14	15130	1780
3	895	172	15	16817	1687
4	1014	119	16	18786	1969
5	1367	353	17	20668	1882
6	1603	236	18	21998	1330
7	2415	812	19	23023	1025
8	3386	971	20	23555	532
9	4980	1594	21	24575	1020
10	7106	2126	22	25096	521
11	9283	2177	23	25431	335
12	11149	1866	24	25950	519

Using the sampled data we find:

H = 12

$\sum_{h=0}^{h=H} V_h$ = 12975

$\sum_{h=0}^{h=H-1} V_h$ = 11149

$\sum_{h=0}^{h=H+1} V_h$ = 13350

Giving a mean value μ_t of 12.83

The standard deviation can be estimated using the normal distribution properties that 68% of all observations fall within one standard deviation of the mean, 95% within two and 99,7% within three. Using these properties and given a total number of messages sent of 25950, we can calculate that:

- 17646 messages are submitted within one standard deviation of the mean
- 24652 within two
- 25872 within three

Using the sampled data we find that 15400 messages are submitted within 4 hours of the mean and 18253 are submitted within 5 hours of the mean; 65% of the messages are submitted sometime in between.

We use a similar equation than the one used to calculate the mean:

$$\sigma = W + \frac{\sum_{w=0}^{w=W} Cv_w - \sum_{w=0}^{w=W-1} Cv_w}{\sum_{w=0}^{w=W+1} Cv_w - \sum_{w=0}^{w=W-1} Cv_w}$$

where: σ is the standard deviation of the distribution
W is the width giving a cumulative total equal to 65% of the grand total
Cv_w is the volume of messages sent for a width of w

Using the sampled data we find:

$$W = 4$$

$$\sum_{w=0}^{w=W} Cv_w = 17646$$

$$\sum_{w=0}^{w=W-1} Cv_w = 15400$$

$$\sum_{w=0}^{w=W+1} Cv_w = 18253$$

Giving a standard deviation of 4.78.

Figure 51 and Figure 52 show the comparison between the measured data and the normal distribution function with the parameters μ and σ. The reader will note that a scale factor has been applied to the normal distribution function so the peak values of both curves match.

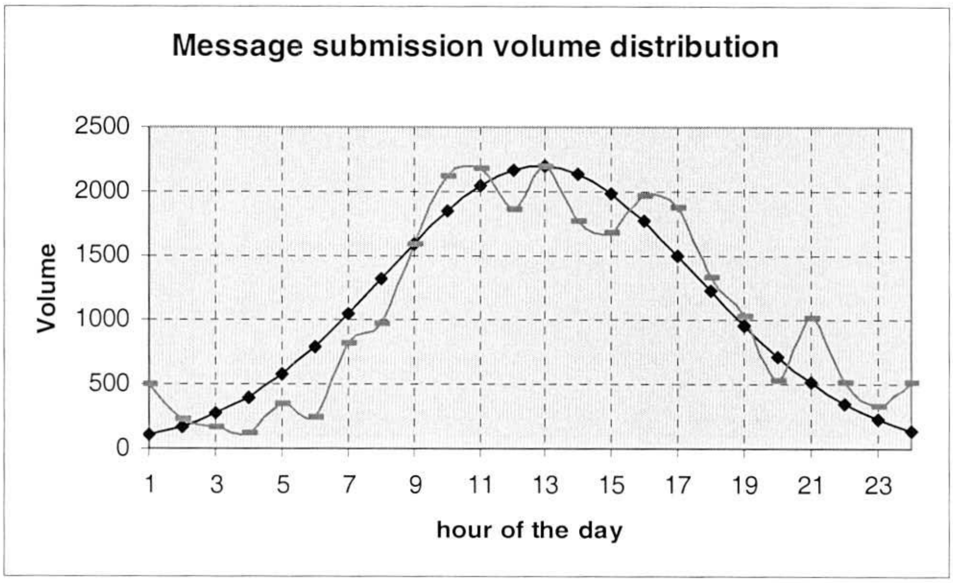


Figure 51: comparison of measured and predicted message volume distribution

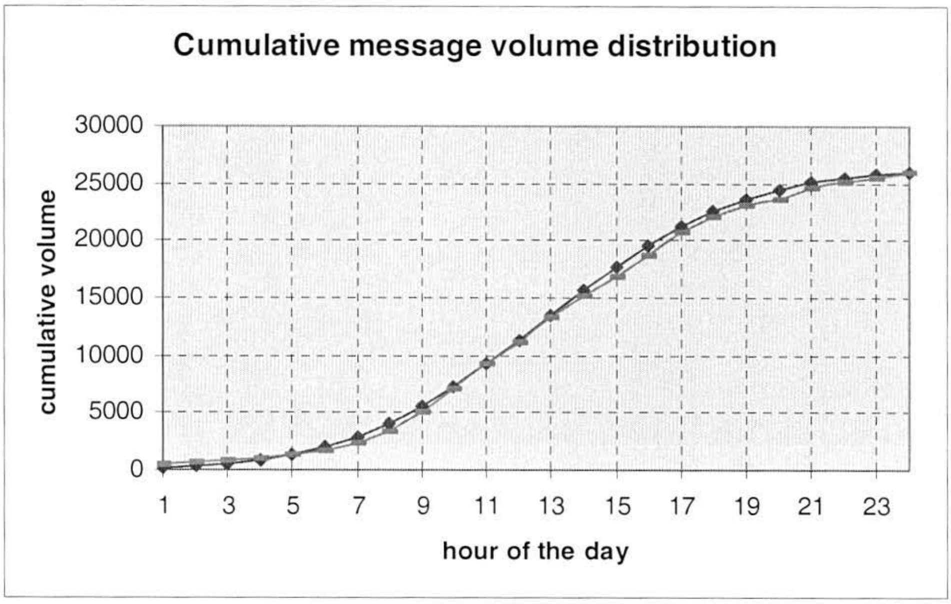


Figure 52: comparison of measured and predicted cumulative message volume distribution

This exercise proves that given the correct sampled data and conditions, one can match the results to known distributions. This is later used to estimate the distributions parameters for the delay in the MTP and the NACP as well as the inter-arrival rates at these layers.

7.3. Test bed description

This section describes the test bed used to evaluate the performance of the gateway (see Figure 53). The live servers are not connected to the same Ethernet hub as the test server in order to avoid access conflicts in the local network segment. Instead, another computer is used as a relay between the live and test servers. It receives e-mails on one interface – network card - and forwards a copy on another interface connected to the test server via the Ethernet hub. This way incoming traffic on the live servers is kept isolated from the test area.

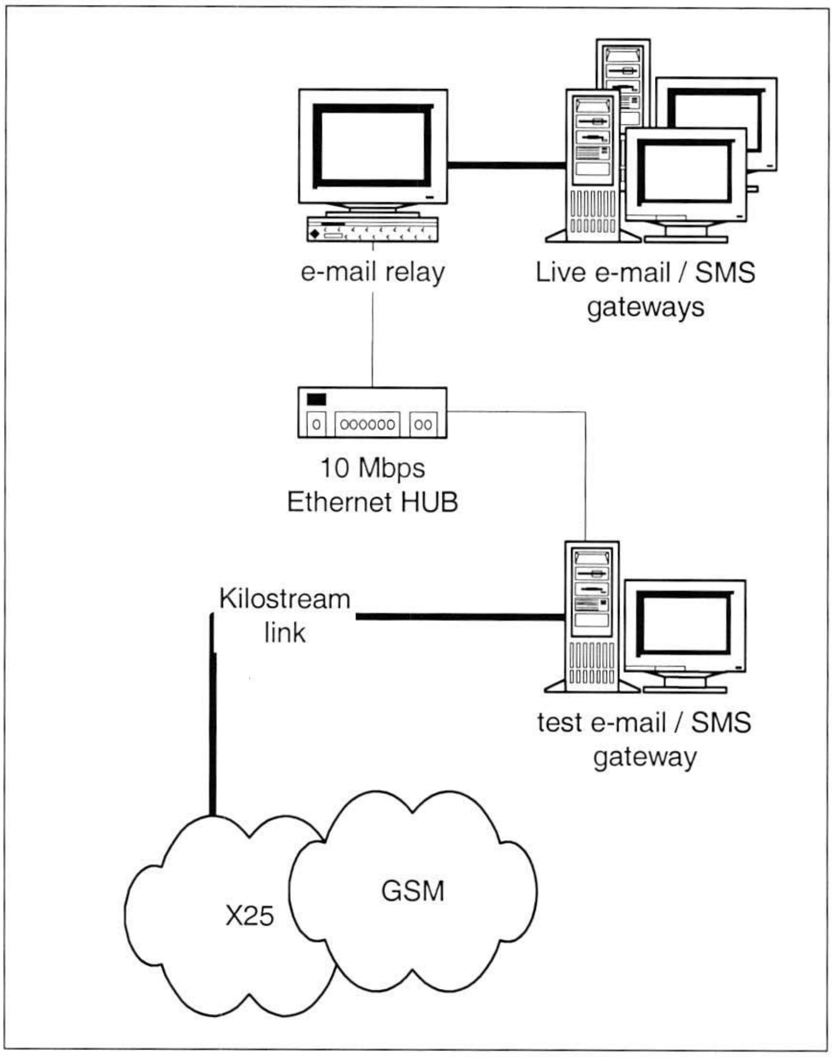


Figure 53: Test bed configuration

The gateway is operating a fairly basic configuration with a single queue for outgoing messages and the routing configured to spool messages for all networks in that queue. The study concentrates on SEMA's SMS2000 submission protocols over an X25 leased line (i.e. kilostream), mainly as it offers the highest achievable throughput. The leased line has a capacity of 64kbps which in theory offers message submission throughput in excess of 30 per second. In practice, the SMSC throttles the submission rate per connection to a more reasonable 4 to 5 per second. Higher throughput is achieved by using multiple queues for outgoing messages and configuring the routing to load share between the queues. Another throughput limiting factor is the RS232 serial interface between the gateway and the X25 terminating equipment which operates at a maximum speed of 19200bps. For the purpose of this study the kilo-stream link will be considered to operate at the speed of the serial interface.

7.4. Measurements

At the end of the period of study, the data logged by the gateway is exported for analysis. A simple script was written to process the data and compute the quantities needed for the rest of the analysis:

- the customer arrival rate at the MTP, or number of emails entering the gateway per unit of time;
- The customer arrival rate at the NACP, or number of messages generated being spooled in the outgoing queue per unit of time;
- The delay at the MTP layer or time needed to process the email
- The waiting time in the outgoing queue, time spent in the queue by a message awaiting submission.
- The delay at the NACP layer or time spent submitting a message

The reader will note that a few assumptions are made:

- The size of the emails received is irrelevant, if enough textual data is extracted from the email then the maximum payload of the message will be used, otherwise the size of the message reflects the size of the textual data in the email;
- Every email received might contain one or more recipient, the delay at the MTP layer however, is computed per recipient and not since the beginning of the life of the process (i.e. the delay to process recipient n will not include the cumulative delays of processing all previous n-1 recipients);
- Concatenation is not used. If the textual data is too long for the maximum payload of a message then it is truncated.
- The gateway used is built to a high enough specification so that the load stays reasonable even for the highest arrival rates. Typically this means that the load never reaches 30%.
- The possibility that a message may require retransmission to the SMSC due to transmission error, or SMSC outages is neglected. In practice this can be detected by excessively high waiting time in the queue, and the corresponding entries in the data under study are discarded.

The exported data consists of 107667 messages sent over a month period; the file size is 24Mb. In order to estimate the arrival rates and delays the month under study is divided into one minute slots and the number of arrivals and the delays aggregated for each slot
The delays introduce by the various layers of the gateway and the corresponding arrival rates are depicted on Figure 54. The inter-arrival rate of short messages on the GSM network ($\lambda_{\text{GSM network}}$) is an aggregation of the departure rate from the NACP layer of the gateway and the background SMS traffic of the network. Close collaboration with the GSM network operator would be necessary in order to estimate the load of the short message service centre (SMSC) at the time of submission as well as the intensity of the traffic in the signalling channel. Network operators are very reluctant to give load and traffic figures away for obvious commercial reasons and thus this study stops when messages are successfully submitted to the SMSC. In any case and under average conditions, the GSM network is in a steady state where messages can be expected to be delivered within a few seconds. High latency in delivering the messages is rare

and usually related to yearly events such as New Year's Eve and Christmas or major SMSC outages.

We define: μ_n (n = 1,2, ... , 4), to be the service rate (processing delay) of the different components involved in the e-mail to short message generation, short message submission and delivery to the GSM terminal. We also define λ_n (n = 1,2, ... , 5), the inter-arrival rate (time elapsed between two e-mails/short messages) at the corresponding components. An important factor in estimating the overall delay in the system is its relation to the length of the messages being generated. As the offered payload of a short message is fairly low (160 7-bit characters), one can guess that a typical email with at least a couple of lines of textual data will fill in the payload fully. Figure 54 shows the distribution of the message length for the duration of the test. As can be expected 90% of the messages have a length of over 150 characters, and consequently, the length of the messages is assumed constant enough not to have implications on the overall delay.

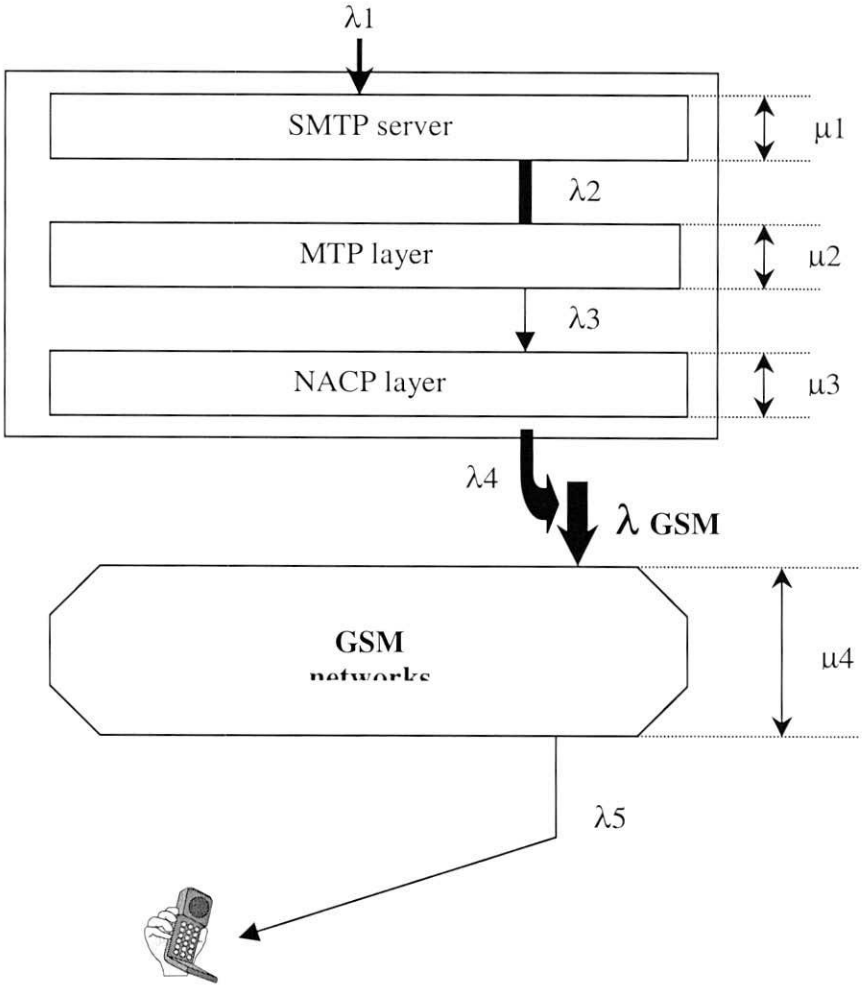


Figure 54: Location of delays and arrival rates in the gateway's component

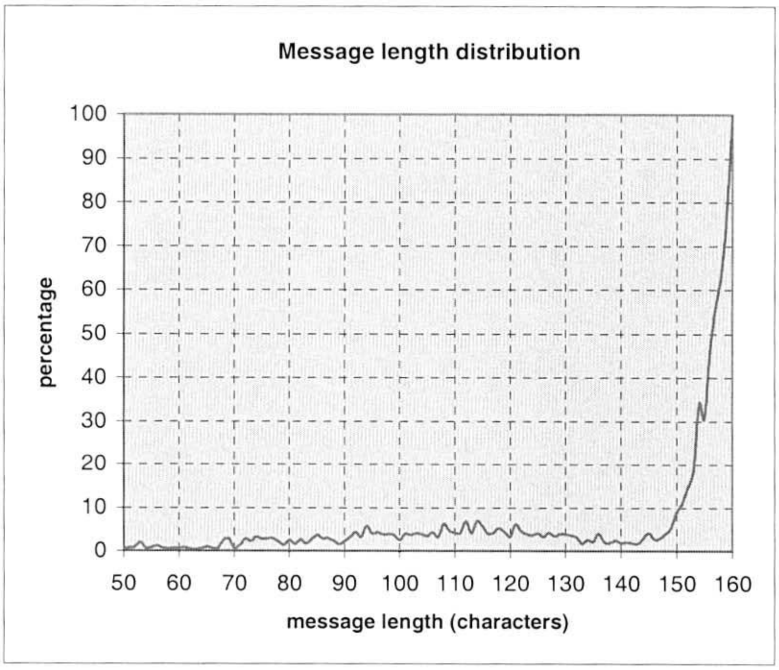


Figure 55: message length distribution

The delay in the MTP layer, NACP layer and the waiting time in the outgoing queue are depicted in Figure 57, Figure 58 and Figure 59. The reader will note that measured delays are highly concentrated in the lower end of the arrival rate. The exponential distribution of the arrival rate on Figure 56 shows indeed that most emails are received at a fairly low rate for the duration of the period.

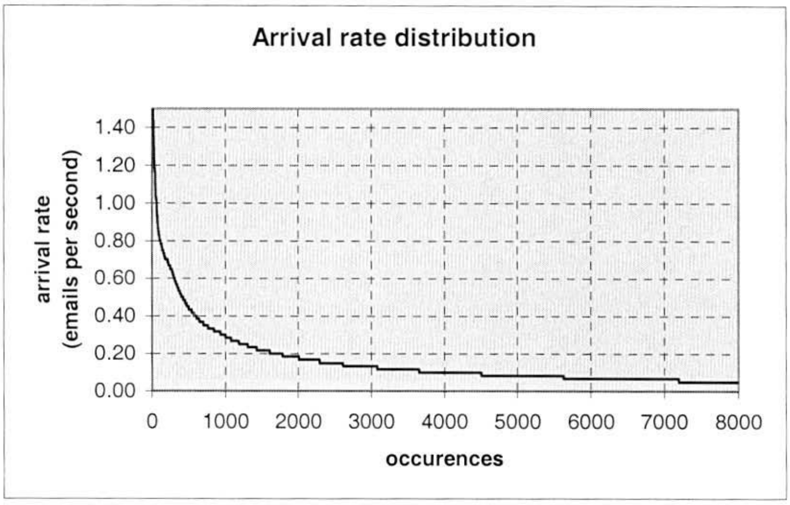


Figure 56: Arrival rate distribution

Another factor to bear in mind is that as traffic is derived from live traffic and the whole test run under real conditions, the expected glitches and abnormal values are found. At the MTP layer this is likely caused by emails with attachments (In extreme cases, it is not unknown for people to try to forward PowerPoint presentation to their phones) which will take considerably more time to perceive and process. At the NACP queue, the waiting time anomalies directly reflect SMSC outages when no message submission is possible.

The trend for relationship between the arrival rate and the delays at each layer is however clearly visible even with the anomalies above mentioned.

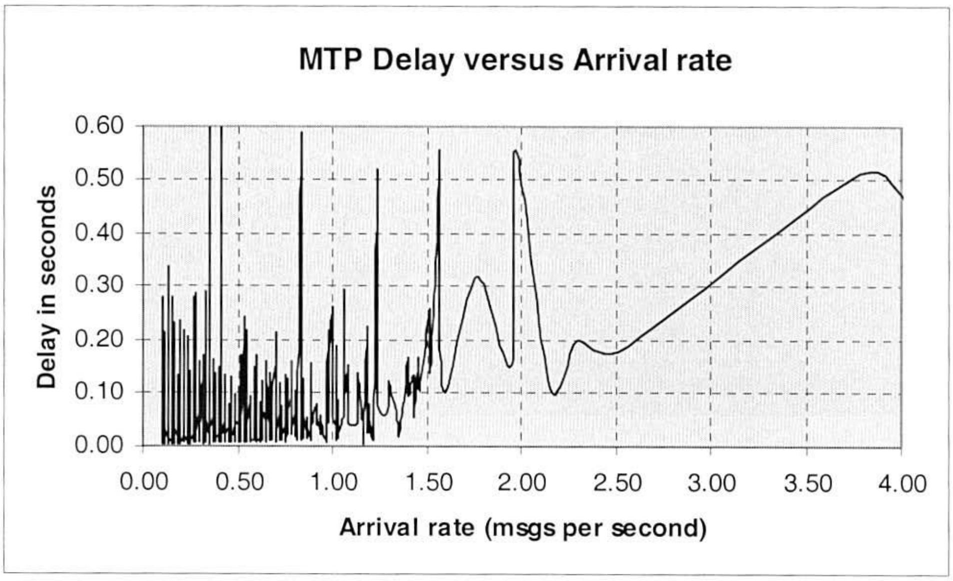


Figure 57: MTP delay versus arrival rate

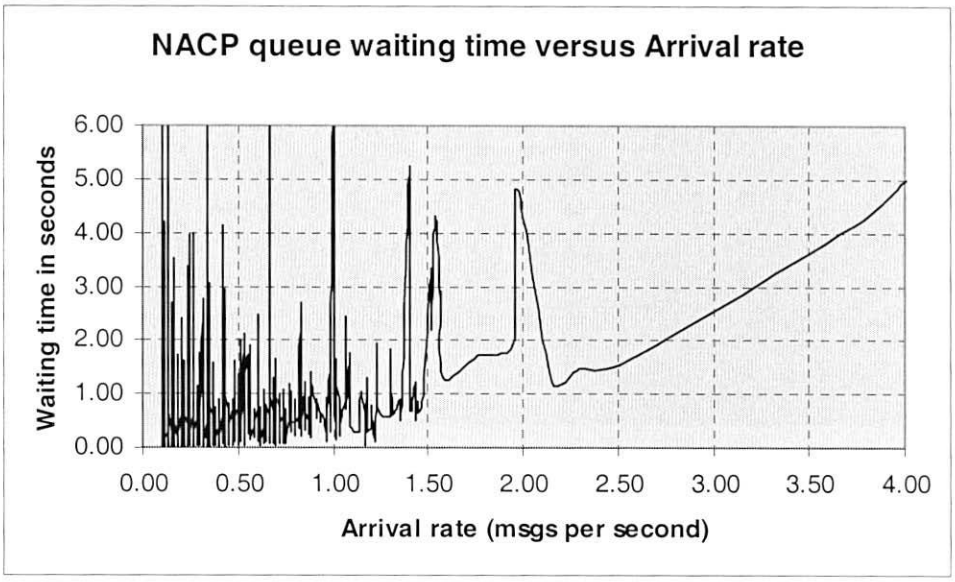


Figure 58: NACP queue waiting time versus arrival rate

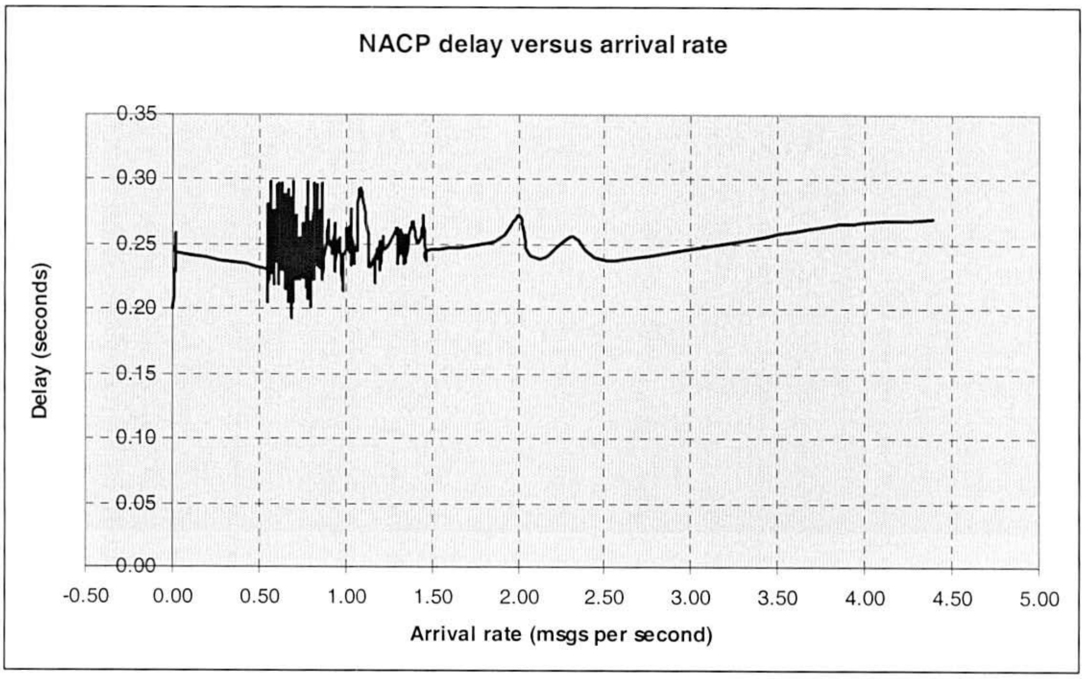


Figure 59: NACP delay versus arrival rate

Figure 59 shows a representation of the NACP delay for a number of arrival rate values. The NACP delay is interesting in that it shows no relationship with the arrival rate. It merely fluctuates between 200 and 300ms and averages 242ms. This is expected as a variation in the arrival rate will have a direct implication on the waiting time in the NACP queue but no effect on the submission time which is dictated by the load on the GSM network's SMSCs. As a consequence, one can therefore estimate that the waiting time W in the queue is a function of the number of messages already queued and the service rate μ_3 at the NACP layer:

$$W = \frac{N_{queued}}{\mu_3}$$

We can then derive the size of the queue at the NACP layer (messages awaiting submission) using this assumption (see Figure 60).

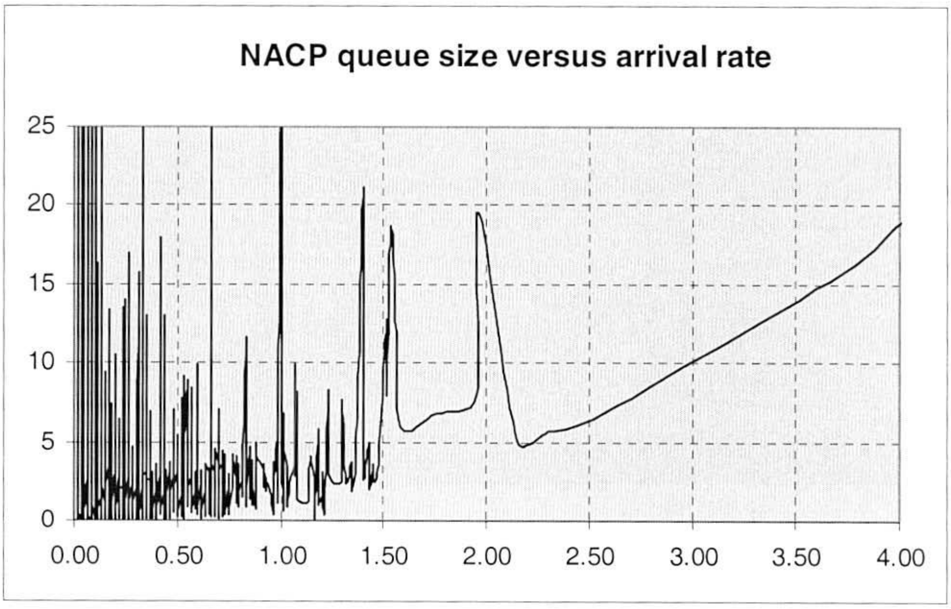


Figure 60: NACP queue size versus arrival rate

The overall delay is obtained by adding the MTP delay, the NACP delay and the waiting time in the NACP queue and is depicted in Figure 61.

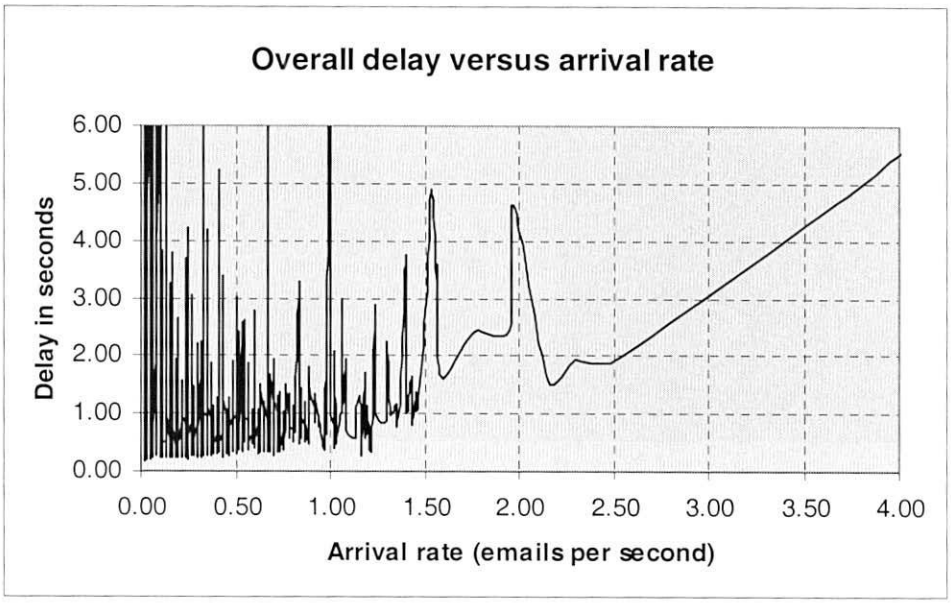


Figure 61: Overall delay versus arrival rate

In the next section, a queueing model is formulated and used to validate the results obtained from the test scenario described in this section.

7.5. Queueing model

The proposed queueing model for the gateway is depicted on Figure 62. The SMTP server layer is not studied as the implementation uses a third party component and no real improvement is possible as long as this is the case. The SMTP server only acts as a receiving node and the delay induced is dominated by the size of the email received as well as a fairly constant start-up time when calling the MTP.

As described earlier in this chapter, each incoming email received by the SMTP server triggers a new instance of the MTP. Providing that the server never reaches a load where resources become sparse, there will always be an instance of the MTP available to parse and process the email. Consequently it was decided to model the MTP layer with an M/M/∞ queue. The distribution of the arrival rate at the MTP layer is indeed exponential as well as the service rate (see Figure 56 and Figure 57). More over, a Poisson process representing the arrival rate in an M/M/∞ queue is generally considered a good model for the aggregate traffic generated by a number of independent users as long as the system being studied operates at relatively mild conditions [Karlin et. al 1975]. Each user contribute to a portion n of the arrival rate λ, and thus itself has an arrival rate of λ/n. In this study, the email traffic is generated by a little more than a thousand independent users.

The NACP layer can be modelled using an M/G/1 queue as the service rate is variable and fluctuates evenly around the mean, and the waiting time in the FIFO queue a function of the number of messages already awaiting submission.

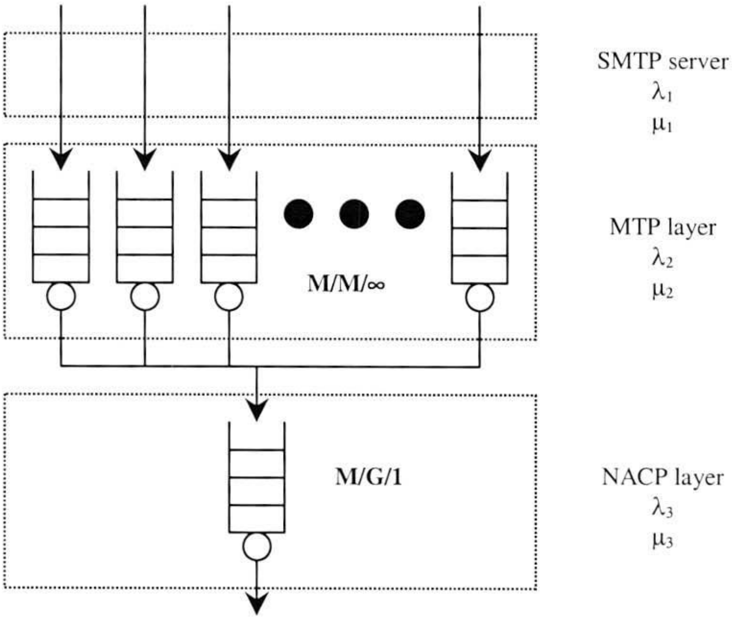


Figure 62: proposed queueing model for the gateway

The system as a whole can be seen as a series of service stations (i.e. in this case processing layers) through which each unit (e.g. email, or message) must travel before being submitted to the SMSC. We have made the assumptions that the aggregated email traffic generated by the customers can be seen as a Poisson process with mean λ. The size of the queues both at the MTP and NACP layer is assumed to be infinite, this is the case as long as the gateway has enough resources to accept incoming mail traffic and the hard disk has enough free space to spool messages. As there are no restrictions on the size of the queue, each layer can be modelled separately.

The departure rate at the MTP layer is the arrival rate at the NACP layer. Imagine a scenario where the size of the queue at the NACP layer was finite and the queue full. The NACP layer would be running at near maximum capacity with the arrival rate nearly exceeding the departure rate. Subsequent emails being processed by the MTP would have to wait for a message to leave the queue before completing the conversion. In this case, the service rate at the MTP layer will reduce dramatically while more and more processes are blocked waiting to spool messages in the NACP queue. Eventually these processes will consume all system's resources and grind the gateway to a halt.

The first version of the gateway used a UNIX FIFO file as the IPC mechanism between the MTP and the NACP. While this approach has the advantage of ease of implementation as read or write access to the file block when the FIFO is empty or full respectively. However, the capacity of the FIFO and in this context also the NACP queue, was finite and fairly small (typically 8Kb or 20 or so messages) and proved too much of a restriction and a huge limiting factor on the scalability of the gateway. More importantly, formulating an analytical model for the system would as a consequence be a lot more difficult. The arrival rate at the NACP queue would have been a function of the current occupation of the queue and the assumptions which are valid for a series of queues would not be possible.

While the current implementation still uses a FIFO as the queueing discipline for the NACP queue, the capacity of the queue is only limited by the available space on the file system. Disk space is fairly good value these days and this is no longer an issue for the purpose of this study.

We now concentrate on the proposed formulated model. The situation is modelled using a two station series. The first station, the MTP layer has no waiting time T_{q2} and the average service time T_{s2} is given by the M/M/∞ delay equation [Kleinrock, L, 1975] as a function of the average service rate μ_2:

$$T_{q2} = 0$$

$$T_{s2} = \frac{1}{\mu_2}$$

This is of course obvious as every arriving customer (e.g. SMTP connection) is granted its own server or instance of the SMTP server. The processing time is then merely the service time which averages $1/\mu_2$.

The second station receives the aggregated departures from all the instances of the MTP layer and functions as an M/G/1 node. The average time in the NACP queue T_{q3} is then given by the Pollaczek-Khinchin (P-K) formula [Kleinrock, L, 1975]:

$$T_{q3} = \frac{\lambda_3^2 \sigma_3^2 + \rho_3}{2\lambda_3 (1 - \rho_3)}$$

Where σ_3 is the variance of the NACP service rate and the utilisation factor of the NACP layer ρ_3 is given by:

$$\rho_3 = \frac{\lambda_3}{\mu_3}$$

Substituting the utilisation factor in the previous equation we have:

$$T_{q3} = \frac{\lambda_3^2\sigma_3^2 + \left(\frac{\lambda_3}{\mu_3}\right)^2}{2\lambda_3\left(1-\left(\frac{\lambda_3}{\mu_3}\right)\right)}$$

$$T_{q3} = \frac{\left(\frac{\lambda_3^2\sigma_3^2\mu_3^2 + \lambda_3^2}{\mu_3^2}\right)}{2\lambda_3\left(\frac{\mu_3 - \lambda_3}{\mu_3}\right)}$$

$$T_{q3} = \left(\frac{\lambda_3^2\sigma_3^2\mu_3^2 + \lambda_3^2}{2\mu_3\lambda_3(\mu_3 - \lambda_3)}\right)$$

$$T_{q3} = \left(\frac{\lambda_3\sigma_3^2\mu_3}{2(\mu_3 - \lambda_3)}\right) + \left(\frac{\lambda_3}{2\mu_3(\mu_3 - \lambda_3)}\right)$$

and finally the average time spent in the NACP layer is equal to T_{q3} plus the average time to submit a message T_{s3},which equals to $1/\mu_3$:

$$T_3 = T_{q3} + T_{s3}$$

and finally:

$$\boxed{T_3 = \left(\frac{\lambda_3\sigma_3^2\mu_3}{2(\mu_3 - \lambda_3)}\right) + \left(\frac{\lambda_3}{2\mu_3(\mu_3 - \lambda_3)}\right) + \frac{1}{\mu_3}}$$

The overall time spent in the gateway is then obtained by adding the time spent in each layer:

$$T = \sum_{n=2}^{n=3} T_{qn} + \sum_{n=2}^{n=3} T_{sn}$$

$$\boxed{T = \frac{1}{\mu_2} + \left(\frac{\lambda_3\sigma_3^2\mu_3}{2(\mu_3 - \lambda_3)}\right) + \left(\frac{\lambda_3}{2\mu_3(\mu_3 - \lambda_3)}\right) + \frac{1}{\mu_3}}$$

Note that the above equation is only valid if the arrival rate does not exceed the service rate of the NACP layer. The values to plot the overall delay can be computed using the measured data.

At the MTP layer the mean and the variance of the service rate are computed for the whole population of the sample. Once again the resulting service rate distribution is normalised so that for an arrival rate of 0 the delay in the system is null.
Figure 63 depicts the plotted values for both the measured delay and the analytical one. As can be expected the overall delay tends to increase dramatically as the system load (i.e. arrival rate over service rate) approaches 100%. Obviously the determining factor in the overall system performance is dictated by the achievable throughput of the connection to the SMSC.
Given a capacity that does not always exceed the requirements of the arrival rate, the NACP queues will build up and in extreme cases (i.e. during a large broadcast of messages) the quality of service offered to the end user will become poor. The queueing discipline of the NACP layer can however benefit from changing from a simple FIFO mechanism to a more elaborate algorithm to still offer reasonable response even if the NACP queue is fairly large. This is discussed in the next chapter.

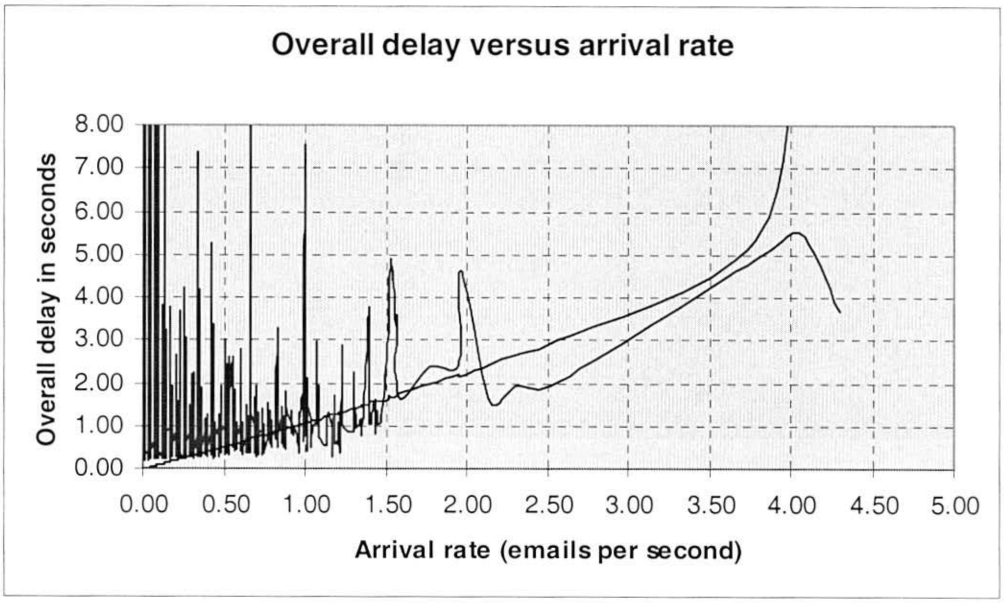


Figure 63: Comparison of the measured and analytical overall delay

7.6. Conclusion

Given the versatility of the gateway and the wide range of scenarii it can be used for, evaluating its performance is not an easy task. However, most commercial implementations will operate the gateway in a fairly basic configuration with a single connection to a network operator. Leased line connectivity to a GSM operator is an expensive exercise running in the thousands of pounds a year and one can argue that multiple links are mainly used for resilience.
This chapter presented a method for evaluating the performance of the gateway in a basic but however common configuration. A decision was made to use part of the live traffic being handled by other gateways, and divert it to an isolated test bed. The rest of the analysis is based on the timings obtained during a one month period. Care was taken to select a user account that would yield a known message volume distribution and cover a wide range of arrival rates. An

analytical model was also formulated to estimate the overall delay in the gateway depending on the arrival rate.

The analytical model, however simplistic, gives an estimation of the overall delay close to the results given by the experiment. As can be expected, most of the delay occurs in the NACP queue when formatted messages are awaiting submission. The capacity of the link and more importantly the speed of the interface to the terminating equipment are the most significant factor and bottleneck. Little can be achieved by optimising the MTP layer in terms of overall performance. However, clever routing configuration and multiple submission interfaces to the network operator would yield higher throughput and offer better scalability and resilience.

In a commercial operating context, when the gateway is used to provide a service to many thousands of users and processes in excess of one hundred thousand messages a day, the NACP queue can grow very large very quickly given even a short outage or slowdown of the SMSC. In this case, in order to maintain a reasonable quality of service, no single point of failure should be allowed and multiple high throughput submission interfaces should be configured.

Given a steady state, where the gateway is constantly busy processing messages, the current FIFO queueing discipline of the NACP offers poor performance and quality of service. A single customer can easily monopolise the NACP layer by submitting a large broadcast of messages to a large population of recipients.

The next chapter addresses the scalability issues of the current implementation and offers an insight on future development and the achievable benefits of changing the queueing discipline of the NACP layer.

8. Directions for further work

This chapter gives guidelines for further work on the implementation of the gateway. The gateway being operated in a production environment, the main issues relate to improving the overall stability and scalability. Some issues, such as the IPC mechanism used, have already been addressed as they impacted reliability at even medium load and had to be dealt with early on. The improved mechanism is now much better suited to high loads and scales nicely. Quality of service can also been improved and a solution is described in this chapter. The proposed alternative has yet to be implemented. Time and resources will dictate how soon this can be achieved.

8.1. Scalability issues

As discussed in the previous chapter, a large contributor to the overall delay in the system is the service rate of the NACP layer. Under most configuration of the gateway a single link to the SMSC will be used and all messages generated will have to leave the system using this link. While in these scenarii, the extra cost of provisioning an extra link is considered too expensive, the queueing discipline of the NACP layer can be altered to make the most of the bandwidth at hand and try to provide a fair share of that bandwidth to every customer.

8.1.1. IPC mechanism optimisation

This issue has been addressed fairly early in the development and evolution of the design of a production gateway used to run a commercial service. Originally UNIX FIFO files were used to act as both the NACP queues and the IPC mechanism. The MTP layer would spool formatted messages in the FIFO and leave it up to the operating system kernel to take care of notifying the NACP layer that one or more messages were ready for submission.While this has the main advantage of not requiring the implementation of a sort algorithm it lacks fairness, security and most of all scalability.

UNIX FIFO have a limited size (typically a few Kb) which means that once the FIFO is full, subsequent attempts to write any messages to it will block until one or more messages have left the queue (e.g. the message data has been read and removed from the FIFO). This would result in a growing number of processes waiting to write to the FIFO, in turn consuming more and more system resources, which can in the worst case consume all the resources of the gateway.

On a day to day basis, this meant that even a short burst of messages could take the gateway down and result in outages which in turn impact customers' confidence in the service. A Replacement solution had to be designed to address the size of the NACP buffer and trigger the appropriate event in the NACP layer.

Of the many IPC mechanisms available in UNIX systems (semaphores, shared memory segment, etc.), UNIX signals seemed like the best compromise in terms of system overhead and ease of implementation. The FIFO file was discarded altogether and the spool directory used to hold all the information on the message spooled. Each message will then generate a specific and unique file in the spool directory and the NACP driver attached to the spool directory (or queue in this context) would be notified of the presence of one or more messages by the means of a UNIX signal. UNIX signals are well suited as a notification mechanism; they do not carry any information but provide a simple mean of notifying a queue that one or more messages are available and await submission. In implementation terms the user defined SIGUSR1 is used for this purpose.

While the previous implementation would have limited capacity NACP queues, the replacement of the IPC mechanism offered unlimited storage for messages (within the file system capacity) and avoided the very damaging gateway outages.

As a result the service rate and delay in the MTP layer decreased dramatically and then most of the delay in processing messages was due to the waiting time in the NACP queue. As messages could be spooled more efficiently and quickly, a new problem appeared. In the previous implementation, each instance of the MTP would be allocated the same priority and time to access the FIFO file (by means of the kernel resources allocation mechanism). In performance terms this meant that each instance of the MTP, whether it had one or a thousand message to spool would be given the same chance of locking the FIFO file and spooling one message at a time before releasing the lock and letting another instance spool a message. Without it being a design goal, this offered crude but almost fair queueing: a large broadcast of messages contained in a single email would not necessarily monopolise the NACP layer and block other messages being submitted.

This led to a new kind of issue altogether, where scalability was achieved to an extent but quality of service was not maintained equally between concurrent users. The queueing discipline of the NACP will have to change in order to achieve more fairness and better response time (on average) through the system.

8.1.2. Towards fair queueing

This section formulates an alternative design for the queueing discipline of the NACP layer. In the current implementation, all messages are also treated equally and are sent in the order in which they were spooled. This is problematic if one customer, knowingly or not, starts to spawn a large number of messages in a short amount of time; typically quicker that the NACP driver can submit. This results in the FIFO filling up mostly if not only with messages belonging to those customers, and most of the outgoing bandwidth to the mobile networks is then used to submit these messages. This scenario is obviously very unfair to concurrent customers whose messages are being delayed significantly; it also offers very little protection over spamming.

The optimum design will have the basic requirements:

- Prevent spamming and monopolisation of an NACP driver for a single customer;
- Provide a mechanism to prioritise messages according to the customer profile;
- Provide a mechanism to allow interactive traffic and higher priority and response time.

The latter is extremely important in the context of 2 way applications such as a quiz game. A group of users is invited to enter a quiz game by the mean of a bulk submission of short messages. As they reply to the messages, the gateway routes the reply to an application server, which in turn submits more messages with the following questions. The response time of the system is essential in this scenario to avoid frustration of the end-user. This kind of traffic should be prioritized in the same way that IP routers handle FTP and Telnet traffic. FTP traffic will have high latency and high throughput while Telnet traffic will benefit from low latency (e.g. good response time) but low throughput. This is the classical problem when a shared medium is used by a number of "greedy" or "selfish" customers.

8.1.3. *Proposed alternate queueing discipline*

The chosen discipline will determine the order in which messages are sent. In an ideal world all customers are treated equally and are expected to be sensible as to the amount of messages they submit at any one time. However, in real life customers are selfish and will gladly take over all the available bandwidth of a server regardless of other customers' needs. While the servers have no mean to control the amount of data that they receive, they can however decide on the order in which the resulting messages are sent.

Imagine a shop with a cash register able to process 3 customers per minute; people wishing to pay join the back of the queue and are served according to the order in which they arrived. This is obviously very unfair if one's shopping trolley contains 1 item while others have several dozen each. The shop then decide to create "less than 10 items" queues and people are able to shop faster. The shop might also let some customers join the queues in the middle as opposed to the end as a reward for always shopping in their store. These customers are assigned a different priority, which also lets them shop faster.

The chosen queueing discipline is inspired from the Stochastic Fairness Queueing (SFQ) model. While it is not quite deterministic in the sense that it does not work in real time; however the model works well for this type of application as long as the queues are sorted at regular intervals. The model is fairly easy to implement and requires relatively little processing power while being almost perfectly fair. SFQ consists of a dynamically allocated number of FIFO queues, one queue per transaction. A transaction can contain one or more messages. Messages are removed from the queues in a round-robin fashion and submitted accordingly, which is why it is called fair. The main advantage of the discipline is that it allows fair sharing of the available bandwidth to the mobile networks amongst customers and prevents one taking over all available resources.

SFQ however, is unable to prioritise message submission according to a customer's profile. This is when Class Based Queueing (CBQ) comes to the rescue. CBQ is a super queue, in that it contains other queues. CBQ can be used in this case to implement priority and further enhance fairness. The super queue contains a number of FIFO queues each corresponding to a certain priority. The queue with the highest priority is emptied first and other queues are not served until all queues with higher priorities are empty.

For a given NACP queue, the queueing discipline sort algorithm is implemented as follows:

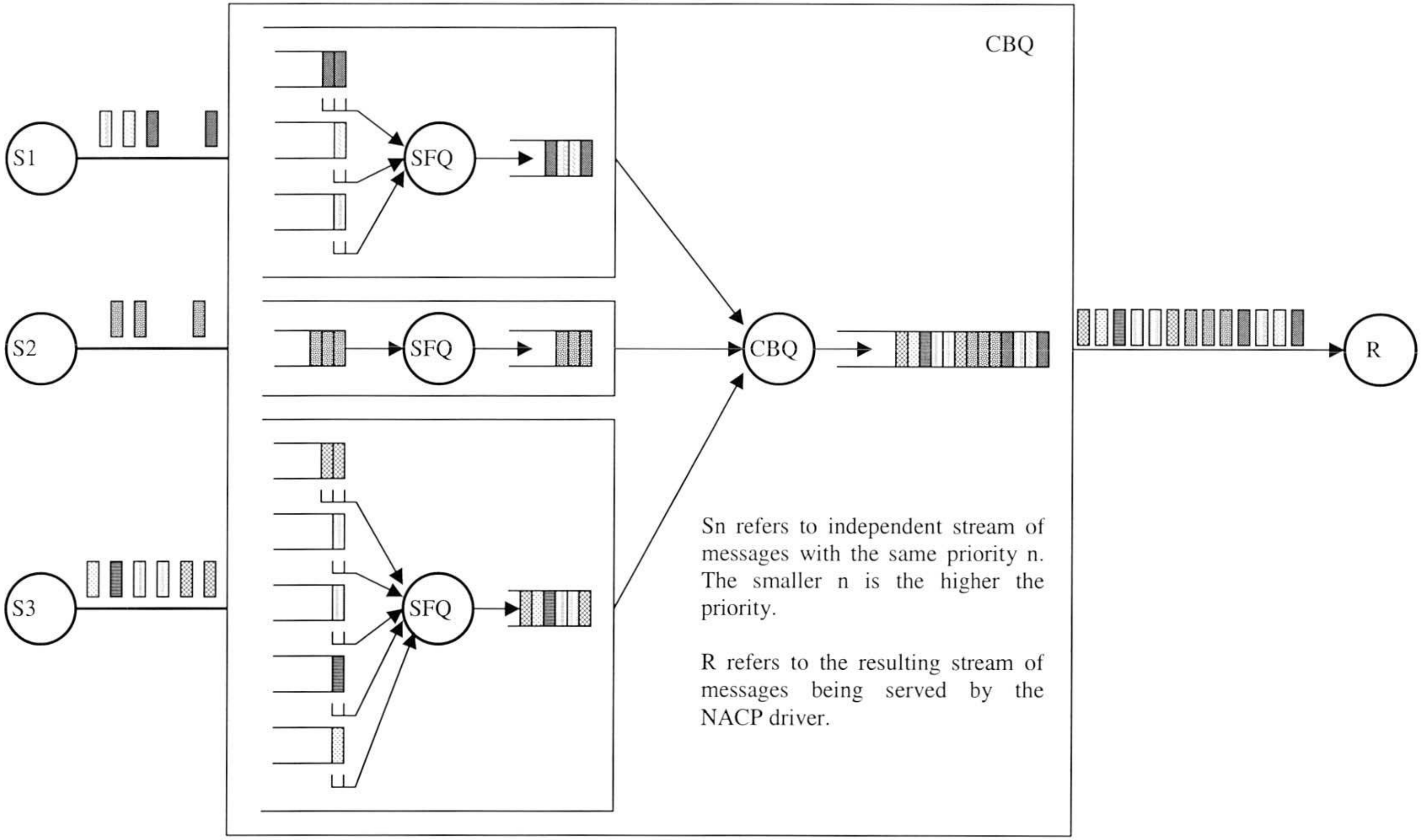
1. Take a snapshot of the queue entries (e.g. read the content of the queue);
2. For each entry corresponding to a message awaiting submission do
    a. Store the filename
3. Sort the snapshot taken
    a. Given 2 filenames to compare sort as follows and in that order:
        i. First sort based on the message priority class;
        ii. Then sort on the message sequence
4. Submit n first messages from the sorted snapshot
5. Go back to step 1

The proposed algorithm is not deterministic as the computing cost would be prohibitively high, it is essential so refresh the snapshot taken and sort it at regular intervals. Ideally, this means that given a flow rate (e.g. NACP service rate) of F_r and a service level S_l (e.g. maximum acceptable time in the system), the snapshot has to be updated at least every M messages sent, with M given by:

$$\boxed{M = F_r S_l}$$

Using this rule, a message being spooled just after the snapshot was taken, will enter the next snapshot within an acceptable delay. This also prevents a large broadcast of messages from overtaking the resources of the NACP.

The implementation follows a three-stage process, in which a CBQ super queue encapsulate a number of dynamically allocated SFQ queues, which are finally served by another CBQ queue.



Sn refers to independent stream of messages with the same priority n. The smaller n is the higher the priority.

R refers to the resulting stream of messages being served by the NACP driver.

Figure 64: Alternate NACP queueing discipline for fairer service

Figure 64 depicts the proposed alternate queueing discipline for the NACP layer. From a simple FIFO model with all its disadvantages, this model addresses the issues highlighted by years of commercial exploitation of the gateway and aims at delivering a better quality of service for the end users.

At the left hand side of the figure, three streams of messages are represented, each associated with a different priority. They enter the main CBQ super queue and are sorted according to their priority. Each stream is then dynamically allocated a FIFO queue in which all messages with the same sequence within a transaction are stored. A transaction will generate message starting with a sequence number of 1 for the first messages up to a sequence number of n for the nth message. The n queues created are then served in a round robin fashion within each CBQ queue and stored in another FIFO awaiting service by the next stage. Finally the n FIFO queues are served by a single CBQ and emptied accordingly.

8.1.4. Practical implementation details

The message store for the NACP drivers is still disk based in this model. A file on disk represents each message physically. The queueing discipline uses the information provided in the filename to identify the priority and sequence number within the transaction of each message. Part of the filenames are used for the queueing discipline but simply ensure uniqueness of the filename within the same NACP queue.

All files have the format tpSSSSYMDhmsRPPPPP with each sub part described in Table 16:

Table 16: alternate queueing discipline file naming convention

Sub part	Description
t	**Message type:** • **c for converted** • **r for retry** • **i for incoming** • **h for deferred** • **t for temporary**
p	**Message priority ranging from 0 highest to 9 lowest**
SSSS	**Sequence number, base62 encoded. Incremented for each message within the same transaction**
Y	**Year, base62 encoded**
M	**Month, base62 encoded**
D	**Day, base62 encoded**
H	**Hour, base62 encoded**
m	**Minute, base62 encoded**
s	**Seconds, base62 encoded**
R	**Random seasoning**
PPPPP	**Process ID of the corresponding MTP instance**

The base62 sequence is 0 ..9 A..Z a..z

The random base62 digit is to help minimise the risk that a new MTP instance could reuse the same PID within the same second. This could lead to a non-unique filename and a loss of data. However, for this to be a problem the new MTP instance would have to spool to the same queue, with the same sequence number within the same second and get the same random number.

Sequence numbers range from (base62) 0000 to zzzz. This represents something like 100 million possible messages within the same transaction.

9. Conclusions

Network congestion and capacity problems resulted in the demand for the design and roll-out of a brand new mobile telephony standard in the early 1980s. Competing analogue mobile communications systems were a plethora and most importantly not compatible. Users were offered very little in terms of valued added services, security, or even roaming possibilities. The "Groupe Spéciale Mobile" was formed to specify and coordinate the development and deployment of the new digital standard for mobile communications, now known as GSM. For the first time in mobile communications history, the coordinated efforts and manufacturers and standard bodies and organisations resulted in the world wide deployment of a fully featured, robust mobile communication technology.

One of the value-added services offered by GSM in phase 2 of its development, was the Short Message Service or SMS. Originally designed as a voicemail notification service, SMS was not supposed to be used on a large scale and consequently used the signalling channel as a bearer. In practice, this meant that the relatively low bandwidth of the signalling channel would be borrowed when idle to transport the text based messaging service. What was possibly a good design as it meant ease of implementation, however meant that should the service become popular, the signalling channel could suffer and quickly become congested.

GSM network users quickly realised that the speed and convenience of SMS made it suitable for much more than voicemail alerts, even if the creation of a text message on a standard handset was tedious to say the least. However, persistent users of SMS were still a small minority and voice services preferred by most users. Sending messages was simply too much hard work and if the service was to gain mass popularity, this process had to be simplified.

Around the same time, many of the biggest network operators in the world started to open access to their message centre and offered a rudimentary dial up service to speed up the submission process. Users could then call an operator or use a simple terminal emulator to type in messages by following a succession of on screen menus. Simple software applications started to appear and would offer a user friendly interface to type and submit short messages by the way of a modem. They almost instantly received consumer's acceptance and a commercial success. Dialogue Communications Ltd, who sponsored this research was founded back in 1995 around the launch of such an application called Pagemail™.

While applications such as Pagemail™ contributed to the popularity of SMS, the user still had to configure and install a modem on his personal computer. In the mid nineties, Internet access was nowhere near as cheap or widely used as it is today and most people did not own a computer for personal use. Access to the Internet and use of its most popular application, electronic mail was still mostly for universities, and large international companies. Nevertheless, a large number of users could be reached and benefit if the right kind of gateway was available.

The design and development of the gateway described in this thesis started back in 1996 from a simple idea. SMS had been designed as a voicemail notification service, and with email becoming more and more popular and people spending more time on the move out of the office, it could also be used as a notification service for emails. On a more general basis, even the short amount of data carried by a short message was suitable for a whole range of notification services, ranging from stocks and shares updates to weather bulletin and even daily horoscope delivery.

The focus on this thesis is on one of the available access mechanisms of the gateway using the Simple Mail Transfer Protocol. The wide availability of client email software meant that the end user would be presented with a familiar interface and message submission would be greatly simplified. The design goals were simple even if the implementation had to hide away the complexity of numerous submission protocols and provide a common API to the developer or researcher.

As usage grew from a couple of hundreds of emails processed every month to a few hundred thousands, evaluating the performance and identifying the bottleneck became a necessity in order to run a successful, scalable and robust service. Other access mechanisms also became necessary as email soon proved too limited and processor intensive for high throughput messaging. Still the same API had to be available regardless of the transport used to carry the payload for the messages.

The work described in this thesis has addressed a gap in the range of tools available to support applications using the short message service as a bearer. The implemented design offers application developers a framework and a toolkit to access the wide variety of parameters supported by the GSM short message service. The wide range of access mechanisms, be it SMTP, HTTP, or XML operating over a common API, hides away the complexity of formatting, routing and submitting short messages, enhancing productivity and yielding faster application development.

More than a phenomenally successful commercial implementation, the gateway architecture is being used daily by research students and development teams. Some projects study the performance and characteristics of the Wireless Application Protocol using SMS as a bearer, others have integrated semantics analysis and text summarisation to make more use of the short payload available. On the commercial side, the gateway offers an easy access to the SMS features and is currently used to implement and evaluate enhanced and future messaging protocols and exciting two way messaging applications.

The development cycle of the gateway is now slowing down and mainly involves bug fixes and minor optimisations. Chapter 8 highlights two areas that would most benefit from improvement:

- The IPC mechanism used and;
- An proposed alternative queueing discipline.

The IPC mechanism has been dealt with and replaced with a simpler signal based one which scales much better and improves the reliability of the gateway at medium and high loads. The queueing discipline has yet to be implemented when time and commercial constraints permit. It is nonetheless an essential evolution towards a fully featured production gateway.

10. References

[Allard et. al. 1994] Allard, J, Sinofsky, S. (1994), 'Getting Wired Into the Internet: A Crash Course on FTP, Gopher, Web, and More,' Microsoft System Journal, Vol. 9, No. 9, Sep. 1994.

[Jan A. Audestad, 1988] Network aspects of the GSM system. In *EUROCON 88*, June 1988.

[D.M. Balston, 1991] The pan-European cellular technology. In R.C.V. Macario, editor, *Personal and Mobile Radio Systems*. Peter Peregrinus, London, 1991.

[D.M. Balston, 1993] The pan-European system: GSM. In D. M. Balston and R.C.V. Macario, editors, *Cellular Radio Systems*. Artech House, Boston, 1993.

[M. Bezler et al, 1993] GSM base station system. *Electrical Communication*, 2nd Quarter 1993.

[Borenstein, N. 1993] Borenstein, N., Freed, N. (1993): 'RFC 1521:- MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies', Sep. 1993.

[CCITT E.164 1997] CCITT E.164 (1997), ' Numbering plan of the international telephone service,' Version 5, 1997.

[CEPT] Conference Européenne des Administration des postes et des telecommunications, online at http://www.cept.org

[David Cheeseman, 1991] The pan-European cellular mobile radio system. In R.C.V. Macario, editor, *Personal and Mobile Radio Systems*. Peter Peregrinus, London, 1991.

[CDPD] http://www.networking.ibm.com/netprod.html

[Cluts 1993] Cluts, N. (1993), 'The ABCs of Converting to Unicode,' Taken from Microsoft Development Library CD 1996, Dec. 1993

[Crocker, D. 1982] Crocker, D., "Standard for the Format of ARPA Internet Text Messages", STD 11, RFC 822, UDEL, Aug. 1982.

[Dechaux *et al.* 1993] Dechaux, C. and Scheller, R. (1993), 'What are GSM and DCS', Electrical Communication, 2nd Quarter, 1993, pp.118-127.

[Dettmer 1997] Dechaux, R, (1997), "Short Message gets longer", IEE Personal Communications journal, May 1997

[ETSI] European Telecommunications Standards Institute, online at http://www.etsi.fr

[ETSI 3.38 1996] ETSI GSM 3.38 (1996) 'Digital cellular telecommunications system (Phase 2+):-Alphabets and language-specific information', European Telecommunications Standards Institute TC SMG, Version 5.2, May. 1996.

[ETSI 3.40 1996] ETSI GSM 3.40 (1996) 'Digital cellular telecommunications system (Phase 2+):-Technical Realisation of the Short Message Service Point-to-Point', European Telecommunications Standards Institute TC SMG, Version 4.13.0, May. 1996.

[ETSI 3.41 1996] ETSI GSM 3.41 (1996): "Digital cellular telecommunication system (Phase 2); Technical realisation of Short Message Service Cell Broadcast (SMSCB)", European Telecommunications Standards Institute TC SMG, Version 5.2.0, May. 1996.

[ETSI 7.05 1996] GSM 07.05 (1996), 'Digital cellular telecommunications system (Phase 2); Use of Data Terminal Equipment - Data Circuit terminating; Equipment (DTE - DCE) interface for Short Message Service (SMS) and Cell Broadcast Service (CBS), European Telecommunications Standards Institute TC SMG, Draft, MAY 1996.

[ETSI 9.07 1996] ETSI GSM 9.07 (1996), 'Digital cellular telecommunications system (Phase 2+); General requirements on interworking between the Public Land Mobile Network (PLMN) and the Integrated Services Digital, Network (ISDN) or Public Switched Telephone Network (PSTN),' European Telecommunications Standards Institute TC.SMG, Version 5.1.0, May 1996.

[ETSI SMG4 1996] ETSI STC SMG4 Tdoc, "a proposed new text for SMS Internet interworking", Sept. 1996.

[Feldmann *et al.* 1993] Feldmann, M. and Rissen, J. P. (1993), 'GSM network systems and Overall System Integration', Electrical Communication, 2nd Quarter, 1993, pp.141-154.

[Freed et. al. 1996] Freed, N. and Borenstein, N. (1996), 'RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies,' available from the website http://www.merseyworld.com/techwatch/ standards/rfcs/rfc2045.html, Nov. 1996.

[Freeman 1994] Freeman, M. (1994), 'Internationalizing Your Win32-based Applications for Windows NT and Windows 95,' Microsoft System Journal, Vol. 9, No. 12, Dec. 1994.

[Freytag, A. 1995] Freytag, A. (1995), 'Build a Multilingual User Interface for Your Application with Win32,' Microsoft System Journal, Vol. 10, No. 4, Apr. 1995

[Gilchrist *et al.* 1994] Gilchrist, P. and Gunther, B. G. (1994), 'General Packet Radio Services on GSM', Mobile Communication International, Spring, 1994, pp. 50-54.

[GPRS] http://www.gsmworld.com/technology/gprs/index.shtml

[J.M. Griffiths, 1992] *ISDN Explained: Worldwide Network and Applications Technology*. John Wiley &Sons, Chichester, 2nd edition, 1992.

[GSMA] GSM Association online at http://www.gsmworld.com

[Haapa *et. al.* 1996] Haapaniema, P. and Hofland, P (1996), 'You Said What?', Byte, Oct. 1996, pp. 48IS 7-10.

[Hall(i) 1994] Hall, W. S. (1994), 'Internationalisation in Windows NT, Part1:- Programming with Unicode,' Microsoft System Journal, Vol. 9, No. 6, Jun. 1994.

[Hall(ii) 1994] Hall, W. S. (1994), 'Internationalization in Windows NT_, Part II: Locales, Languages, and Resources,' Microsoft System Journal, Vol. 9, No. 7, Jul. 1994.

[HSCSD] http://www.gsmworld.com/technology/hscsd/index.shtml

[I. Harris, 1993] Data in the GSM cellular network. In D. M. Balston and R.C.V. Macario, editors, *Cellular Radio Systems*. Artech House, Boston, 1993.

[I. Harris, 1993] Facsimile over cellular radio. In D. M. Balston and R.C.V. Macario, editors, *Cellular Radio Systems*. Artech House, Boston, 1993.

[Thomas Haug. 1988] Overview of the GSM project. In *EUROCON 88*, June 1988.

[H. Lobensommer et. al. 1992] GSM - a European mobile radio standard for the world market. *Telcom Report International*, 15(3-4), 1992.

[Jayapalan et. al. 1994] Jayapalan, J. and Burke, M. (1994), 'Cellular Data Services Architecture and signalling,' IEEE Personal Communications, 2nd Quarter, 1994, pp. 44-55.

[Karlin et. al 1975] Karlin, S., and Taylor, H. M. 1975. A first course in stochastic processes. New York: Academic Press.

[Kano(i) 1995] Kano, N. (1995), 'Developing International Software for Windows 95 and Windows NT', Microsoft Press, July 1995, ISBN: 1-55615-840-8.

[Kano(ii) 1995] Kano, N. (1995), 'Putting on international interface, 'Tips On Designing UIs That Can Be Accepted Around the World,' Microsoft Developer Network News, Vol. 4, No. 2, Mar. 1995.

[Kano(iii) 1995] Kano, N. (1995), 'Common IME System on Far East Windows 95', Microsoft Developer Network News, Vol. 4, No. 5, Sep. 1995.

[Kano *et al.* 1994] Kano, N. and Freytag, A. (1994), 'The International Character Set Conundrum: ANSI, Unicode, and Microsoft Windows', Microsoft System Journal, Vol. 9 (11). Also from the 'Microsoft development library CD 14' Jan. 1996.

[Kleinrock, L, 1975]. 'Queueing Systems, Vol. 1. New York: Wiley.

[Kleinrock, L, 1976] 'Queueing Systems, Vol. 2. New York: Wiley.

[Kuhn *et al.* 1994] Kuhn, P. J., Pack, C. D., Skoog, R. G. (1994), 'Common Channel Signalling Networks: Past, Present, Future,' IEEE Journal on Selected Area in Communication, Vol. 12 (3), Apr. 1994, pp.381-194.

[Leisher 1996] Leisher, M. (1996), 'An adventure in implementing Unicode support on UNIX platform,' 9th International Unicode Conference, San Jose, CA, Sep. 1996, pp. 1-8.

[B.J.T. Mallinder, 1988] Specification methodology applied to the GSM system. In *EUROCON 88*, June 1988.

[Mobeen *et al.* 1995] Mobeen, K, and Kilpatrick, J. (1995) "Mobitex and mobile data standards", IEEE Communications magazine, March 1995, pp96-101.

[MOBITEX] http://www.ericsson.com/network_operators/mobitex/about.shtml

[Mouly *et al.* 1995] Mouly, M., and Pautet, M. B. (1995) 'Current Evolution of the GSM Systems,' IEEE Personal Communications, Vol. 2 (5), Oct. 1995, pp.9-19.

[Myers(i) 1995] Myers, S. (1995), 'Japanese Character Sets and Encoding Methods for PC,' Computing Japan Magazine, http://www.cjmag.co.jp/, Nov. 1995.

[Myers(ii) 1995] Myers, S. (1995), 'A Unicode Tutorial,' Computing Japan Magazine, http://www.cjmag.co.jp/, Dec. 1995.

[J.E. Natvig et.al. 1989] Speech processing in the pan-European digital mobile radio system (GSM) - system overview. In *IEEE GLOBECOM 1989*, November 1989.

[Perkins 1996] Perkins, K. (1996), "RFC 2003:- IP Encapsulation within IP", available from the website http://www.merseyworld.com/techwatch/standards/rfcs/rfc2003.html ,October 1996

[Postel J.B, 1982] Postel J.B "Simple Mail Transfer Protocol", RFC 821, 1982

[Q.700] CCITT Recommendation Q.701: Introduction to CCITT Signalling System No.7, FASCICLE V1.7, Nov. 1988.

[Q.701] CCITT Recommendation Q.701: Functional Description of the Message Transfer Part Signalling System No.7, FASCICLE V1.7, Nov. 1988.

[Q.711] CCITT Recommendation Q.711: Functional Description of the Signalling Connection Control Part, FASCICLE V1.7, Nov. 1988.

[Q.771] CCITT Recommendation Q.771: Description of Transaction Capabilities of Signalling System No.7, FASCICLE V1.7, Nov. 1988.

[Rahnema 1993] Rahnema, M.(1993), 'Overview of the GSM system and protocol architecture', IEEE Communication Magazine, Vol. 3 (4), Apr. 1993, pp.92-100.

[RD-LAP] http://www.motorola.com/LMPS/RNSG/data/networks/

[RFC 2110] Palme, J., Hopmann, A. (1997), 'RFC 2110:- MIME E-mail Encapsulation of Aggregate Documents, such as HTML', Mar. 1997.

[Roth 1993] Roth, W. (1993), 'Data service on the GSM platform,' GSM Summit Hong Kong, March 1993.

[Scourias 1994] John Scourias, A brief overview of GSM, University of Waterloo

[S. Mohan et. Al. 1994] Two user location strategies for personal communication services. *IEEE Personal Communications*, 1(1), 1994.

[C. B. Southcott et al, 1989] Voice control of the pan-European digital mobile radio system. In *IEEE GLOBECOM 1989*, November 1989

[Spaniol *et al.* 1995] Spaniol, O., Fasbender, A., Hoff, S., Kaltwasser, J., Kassubek, J.(1995) 'Impacts of Mobility on Telecommunication and Data Communication Networks', IEEE Personal Communications, Vol. 2 (5), Oct. 1995, pp.20-33.

[TSCR 1995] Telecom Securicor Cellular Radio ltd (1995), "Interface Specification for the connection of external systems to the Cellnet GSM Short Message Service centre using the Telecator Alphanumeric Protocol for mobile terminated short messaging".

[P. Vary et al. 1989] Speech codec for the European mobile radio system. In *IEEE GLOBECOM 1989*, November 1989.

[Varin *et al.* 1993] Varin, J., Bezler, M. Hofmans, R. and Van Den Bossche, K. (1993), 'GSM Base Station System', Electrical Communication, 2nd Quarter, 1993, pp.155-163.

[Vu *et. al.* 1995] Vu, C. Satoh, S. and Grove, M. (1995), 'Multibyte Character Set (MBCS) Survival Guide,' Taken from Microsoft Development Library CD 1996, Aug. 1995

[WAP] http://www.gsmworld.com/technology/wap/index.shtml

[C. Watson. 1993] Radio equipment for GSM. In D. M. Balston and R.C.V. Macario, editors, *Cellular Radio Systems*. Artech House, Boston, 1993.

11. Appendix A: Publications produced in the course of this study

G. Peersman, S. Cvetkovic, C. Smythe, P. Griffiths, and H. Spear, "The Global System for Mobile Communications Short Message Service", **IEEE Personal Communication Magazine**, June 2000 issue, Vol. 7, No. 3, pp 15-23.

G. Peersman, S. Cvetkovic, C. Smythe, P. Griffiths, and H. Spear, "A Tutorial Overview of the Short Message Service within GSM", **IEE Computing & Control Engineering Journal**, April 2000, Vol. 11 No. 2, pp 79-89.

K. Koumpis, S. Cvetkovic and G. Peersman, "Performance Evaluation of SMS-Based Email and Voicemail Notification Architecture", **Proc. of the 5th Workshop on Emerging Technologies in Telecommunications**, pp. 282-286, Bayona, Spain, Sept. 1999.

S Gallagher, J Moore, S Cvetkovic, C Smythe, I Stergiou and G Peersman, "Generic Architecture for Information Availability(GAIA) Infrastructure Statement", July 1997, **AC211 WP10 1001, Ref. AC221/NW/WP10/DS/P/010/b1**.

G.Peersman, S. Cvetkovic, C. Smythe, P.Gritffiths, and H.Spear, "The integration of SMS with voice based technology", **IEE Colloquium on Advances in Interactive Voice Technologies for Telecommunications Services**, June 1997, pp 9/1-9/7